



Konstantin K. Likharev
Essential Graduate Physics
Lecture Notes and Problems

Beta version

Open online access at
<https://sites.google.com/site/likharevegp/>
and
<http://commons.library.stonybrook.edu/egp/>

Part EM:

Classical Electrodynamics

Table of Contents

Chapter 1. Electric Charge Interaction (20 pp.)

- 1.1. The Coulomb law
- 1.2. The Gauss law
- 1.3. Scalar potential and electric field energy
- 1.4. Exercise problems (17)

Chapter 2. Charges and Conductors (64 pp.)

- 2.1. Electric field screening
- 2.2. Capacitance
- 2.3. The simplest boundary problems
- 2.4. Other orthogonal coordinates
- 2.5. Variable separation
- 2.6. Charge images
- 2.7. Green's functions
- 2.8. Numerical methods
- 2.9. Exercise problems (33)

Chapter 3. Polarization of Dielectrics (26 pp.)

- 3.1. Electric dipole
- 3.2. Dipole media
- 3.3. Linear dielectrics
- 3.4. Molecular field effects
- 3.5. Electric field energy in a dielectric
- 3.6. Exercise problems (20)

Chapter 4. DC Currents (14 pp.)

- 4.1. Continuity equation and the Kirchhoff laws
- 4.2. The Ohm law
- 4.3. Boundary problems
- 4.4. Dissipation power
- 4.5. Exercise problems (9)

Chapter 5. Magnetism (42 pp.)

- 5.1. Magnetic interaction of currents
- 5.2. Vector-potential and the Ampère law
- 5.3. Magnetic flux, energy, and inductance
- 5.4. Magnetic dipole moment, and magnetic dipole media
- 5.5. Magnetic materials
- 5.6. Systems with magnetics
- 5.7. Exercise problems (23)

Chapter 6. Time-Dependent Electromagnetism (32 pp.)

- 6.1. Electromagnetic induction
- 6.2. Quasistatic approximation and skin effect
- 6.3. Electrodynamics of superconductivity and gauge invariance
- 6.4. Electrodynamics of macroscopic quantum phenomena
- 6.5. Inductors, transformers, and ac Kirchhoff laws
- 6.6. Displacement currents
- 6.7. Finally, the full Maxwell equation system
- 6.8. Exercise problems (22)

Chapter 7. Electromagnetic Wave Propagation (66 pp.)

- 7.1. Plane waves
- 7.2. Attenuation and dispersion
- 7.3. Kramers-Kronig relations
- 7.4. Reflection
- 7.5. Refraction
- 7.6. Transmission lines: TEM waves
- 7.7. H and E waves in metallic waveguides
- 7.8. Dielectric waveguides and optical fibers
- 7.9. Resonators
- 7.10. Energy loss effects
- 7.11. Exercise problems (30)

Chapter 8. Radiation, Scattering, Interference, and Diffraction (36 pp.)

- 8.1. Retarded potentials
- 8.2. Electric dipole radiation
- 8.3. Wave scattering
- 8.4. Interference and diffraction
- 8.5. The Huygens principle
- 8.6. Diffraction on a slit
- 8.7. Geometrical optics placeholder
- 8.8. Fraunhofer diffraction from more complex scatterers
- 8.9. Magnetic dipole and electric quadrupole radiation
- 8.10. Exercise problems (20)

Chapter 9. Special Relativity (54 pp.)

- 9.1. Einstein postulates and the Lorentz transform
- 9.2. Relativistic kinematic effects
- 9.3. 4-vectors, momentum, mass, and energy
- 9.4. More on 4-vectors and 4-tensors
- 9.5. Maxwell equations in the 4-form
- 9.6. Relativistic particles in electric and magnetic fields
- 9.7. Analytical mechanics of charged particles
- 9.8. Analytical mechanics of electromagnetic field
- 9.9. Exercise problems (34)

Chapter 10. Radiation by Relativistic Charges (38 pp.)

- 10.1. Liénard-Wiechert potentials
- 10.2. Radiation power
- 10.3. Synchrotron radiation
- 10.4. Bremsstrahlung and Coulomb losses
- 10.5. Density effects and the Cherenkov radiation
- 10.6. Radiation's back-action
- 10.7. Exercise problems (12)

* * *

Additional files (available upon request):

Exercise and Test Problems with Model Solutions ($220 + 69 = 289$ problems; 367 pp.)

Chapter 1. Electric Charge Interaction

This brief chapter describes the basics of electrostatics, the study of interactions between static (or slowly moving) electric charges. Much of this material should be known to the reader from his or her undergraduate studies; because of that, the explanations will be very succinct.¹

1.1. The Coulomb law

A serious discussion of the Coulomb law² requires a common agreement on the meaning of the following notions:³

- *electric charges* q_k , as revealed, most explicitly, by experimental observation of *electrostatic interaction* between the charged particles;
- *electric charge conservation*, meaning that the algebraic sum of q_k of all particles inside any closed volume is conserved, unless the charged particles cross the volume's border; and
- a *point charge*, meaning the charge of an ultimately small ("point") particle whose position in space may be completely described (in a given reference frame) by its radius-vector $\mathbf{r} = \mathbf{n}_1 r_1 + \mathbf{n}_2 r_2 + \mathbf{n}_3 r_3$, where \mathbf{n}_j (with $j = 1, 2, 3$) are unit vectors directed along 3 mutually perpendicular axes, and r_j are the corresponding Cartesian components of \mathbf{r} .

I will assume that these notions are well known to the reader - though my strong advice is to give some thought to their vital importance. Using them, the *Coulomb law* for the electrostatic interaction of two point charges in otherwise free space may be formulated as follows:

$$\mathbf{F}_{kk'} = \kappa q_k q_{k'} \frac{\mathbf{r}_k - \mathbf{r}_{k'}}{|\mathbf{r}_k - \mathbf{r}_{k'}|^3}, \quad (1.1)$$

Coulomb
law for
2 point
charges

where $\mathbf{F}_{kk'}$ denotes the force exerted on charge number k by charge number k' . This law is certainly very familiar to the reader, but several remarks may still be due:

(i) Flipping indices k and k' , we see that Eq. (1)⁴ complies with the 3rd Newton law: the reciprocal force is equal in magnitude but opposite in direction: $\mathbf{F}_{k'k} = -\mathbf{F}_{kk'}$.

(ii) According to Eq. (1), the magnitude of the force, $F_{kk'}$, is inversely proportional to the square of the distance between the two charges – the well-known undergraduate-level formulation of the Coulomb law.

¹ For remedial reading, virtually any undergraduate text on electricity and magnetism may be used; I can recommend either the classical text by I. Tamm, *Fundamentals of Theory of Electricity*, Mir, 1979, or the more readily available textbook by D. Griffiths, *Introduction to Electrodynamics*, 3rd ed., Prentice-Hall, 1999.

² Discovered experimentally in the early 1780s, and formulated in 1785 by C.-A. de Coulomb.

³ On the top of the more general notions of *classical Cartesian space*, *point particles* and *forces*, which are used in classical mechanics – see, e.g., CM Sec. 1.1. (Acronyms CM, SM, and QM refer to other three parts of my lecture note series. In those parts, this Classical Electrodynamics part is referred to as EM.)

⁴ As in all other parts of my lecture notes, chapter numbers are omitted in references to equations, figures, and sections within the same chapter.

(iii) Since vector $(\mathbf{r}_k - \mathbf{r}_{k'})$ is directed from point $\mathbf{r}_{k'}$ toward point \mathbf{r}_k (Fig. 1), Eq. (1) implies that charges of the same sign (i.e. with $q_k q_{k'} > 0$) repulse, while those with opposite signs ($q_k q_{k'} < 0$) attract each other.

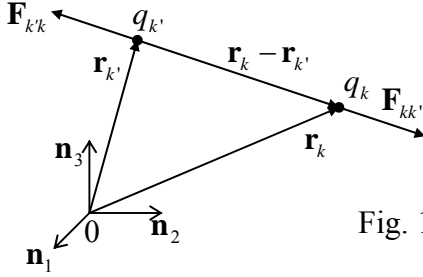


Fig. 1.1. Direction of the Coulomb forces (for $q_k q_{k'} > 0$).

(iv) Constant κ in Eq. (1) depends on the system of units we use. In the *Gaussian* units, κ is set to 1, for the price of introducing a special unit of charge (the *statcoulomb*) that would fit the experimental data for Eq. (1), for forces $\mathbf{F}_{kk'}$ measured in Gaussian units (*dynes*). On the other hand, in the *International System* (“SI”) of units, the charge unit is one *coulomb* (abbreviated C),⁵ close to 3×10^9 statcoulombs, and κ is different from unity:⁶

$$\kappa_{\text{SI}} = \frac{1}{4\pi\epsilon_0} \equiv 10^{-7} c^2. \quad (1.2)$$

Electric
force
constant

Unfortunately, the continuing struggle between zealot proponents of these two systems bears all ugly features of a religious war, with a similarly slim chances for any side to win it in any foreseeable future. In my humble view, each of these systems has its advantages and handicaps (to be noted on several occasions below), and every educated physicist should have no problem with using any of them. Following insisting recommendations of international scientific unions, I will mostly use SI units, but for readers’ convenience, duplicate the most important formulas in the Gaussian units.

Besides Eq. (1), another key experimental law of electrostatics is the *linear superposition principle*: the electrostatic forces exerted on some point charge (say, q_k) by other charges do not affect each other and add up as vectors to form the net force:

$$\mathbf{F}_k = \sum_{k' \neq k} \mathbf{F}_{kk'}, \quad (1.3)$$

where the summation is extended over all charges but q_k , and the partial force $\mathbf{F}_{kk'}$ is described by Eq. (1).⁷ The fact that the sum is restricted to $k' \neq k$ means that a *point charge does not interact with itself*.

⁵ In the formal metrology, one coulomb is defined as the charge carried over by a constant current of one ampere (see Ch. 5 for its definition) during one second.

⁶ Constant ϵ_0 is called either the *electric constant* or the *free space permittivity*; from Eq. (2) with the free-space speed of light $c \approx 3 \times 10^8$ m/c, $\epsilon_0 \approx 8.85 \times 10^{-12}$ SI units. For more accurate values of the constants, and their brief discussion, see appendix CA: *Selected Physical Constants*.

⁷ Physically this is a very strong statement: it means that Eq. (1) is valid for any pair of charges regardless of presence of other charges, i.e. not only in the free space, but in also placed into an arbitrary medium. The apparent modification of this relation by conductors (Ch. 2) and dielectrics (Ch. 3) is just the result of appearance of additional electric charges within those media.

This fact may look trivial from Eq. (1), whose right-hand part diverges at $\mathbf{r}_k \rightarrow \mathbf{r}_{k'}$, but becomes less evident (though still true) in quantum mechanics where the charge of even an elementary particle is effectively spread around some volume, together with particle's wavefunction.⁸

Now we may combine Eqs. (1) and (3) to get the following expression for the net force \mathbf{F} acting on some charge q located at point \mathbf{r} :

$$\mathbf{F} = q \frac{1}{4\pi\epsilon_0} \sum_{\mathbf{r}_k' \neq \mathbf{r}} q_{k'} \frac{\mathbf{r} - \mathbf{r}_{k'}}{|\mathbf{r} - \mathbf{r}_{k'}|^3}. \quad (1.4)$$

This equation implies that it makes sense to introduce the notion of the *electric field* at point \mathbf{r} , as an entity independent of the *probe charge* q , characterized by vector

$$\mathbf{E}(\mathbf{r}) \equiv \frac{\mathbf{F}}{q}, \quad (1.5)$$

formally called the *electric field strength* – but much more frequently, just the “electric field”. In these terms, Eq. (4) becomes

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \sum_{\mathbf{r}_k' \neq \mathbf{r}} q_{k'} \frac{\mathbf{r} - \mathbf{r}_{k'}}{|\mathbf{r} - \mathbf{r}_{k'}|^3}. \quad (1.6)$$

This concept is so appealing that Eq. (5) is used well beyond the boundaries of free-space electrostatics. Moreover, the notion of field becomes virtually unavoidable for description of time-dependent phenomena (such as electromagnetic waves), where the electromagnetic field shows up as a specific form of matter, with zero rest mass, and hence different from the usual “material” particles.

Many problems involve many point charges $q_{k'}$, $q_{k''}$, ..., located so closely that it is possible to approximate them with a *continuous charge distribution*. Indeed, for a group of charges within a very small volume d^3r' , with the linear size satisfying strong condition $dr \ll |\mathbf{r}_k - \mathbf{r}_{k'}|$, the geometrical factor in Eq. (6) is essentially the same. As a result, all these charges may be treated as a single charge $dQ(\mathbf{r}')$. Since this charge is proportional to d^3r' , we can define the local (3D) *charge density* $\rho(\mathbf{r}')$ by relation⁹

$$\rho(\mathbf{r}') d^3r' \equiv dQ(\mathbf{r}') \equiv \sum_{\mathbf{r}_{k'} \in d^3r'} q_{k'}, \quad (1.7)$$

and rewrite Eq. (6) as

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \sum_{d^3r'} dQ(\mathbf{r}') \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} = \frac{1}{4\pi\epsilon_0} \sum_{d^3r'} \rho(\mathbf{r}') d^3r' \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3}, \quad (1.8)$$

⁸ Moreover, there are some widely used approximations, e.g., the Kohn-Sham equations in the density functional theory of multiparticle systems, which essentially violate this law, thus limiting the accuracy and applicability of these approximations - see, e.g., QM Sec. 8.4.

⁹ The 2D (areal) charge density σ and 1D (linear) density λ may be defined absolutely similarly: $dQ = \sigma d^2r$, $dQ = \lambda dr$. Note that a finite value of σ and λ means that the volume density ρ is infinite in the charge location points; for example for a plane $z = 0$, charged with a constant areal density σ , $\rho = \sigma \delta(z)$.

i.e. as the integral (over the whole volume containing all essential charges):¹⁰

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int \rho(\mathbf{r}') \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} d^3 r'. \quad (1.9)$$

Coulomb
law for
continuous
charge
distribution

It is very convenient that Eq. (9) may be used even in the case of discrete point charges, employing the notion of Dirac's δ -function,¹¹ which is a mathematical approximation for a very sharp function equal to zero everywhere but one point, and still having a finite (unit) integral. Indeed, in this formalism, a set of point charges $q_{k'}$ located in points $\mathbf{r}_{k'}$ may be presented by the pseudo-continuous distribution with density

$$\rho(\mathbf{r}') = \sum_{k'} q_{k'} \delta(\mathbf{r}' - \mathbf{r}_{k'}). \quad (1.10)$$

Plugging this expression into Eq. (9), we come back to the discrete version (6) of the Coulomb law.

1.2. The Gauss law

Due to the extension to point ("discrete") charges, it may seem that Eqs. (5) and (9) is all we need for solving any problem of electrostatics. In practice, this is not quite true, first of all because the direct use of Eq. (9) frequently leads to complex calculations. Indeed, let us consider a very simple example: the electric field produced by a spherically-symmetric charge distribution with density $\rho(r')$. We may immediately use the problem symmetry to argue that the electric field should be also spherically-symmetric, with only one component in spherical coordinates: $\mathbf{E}(\mathbf{r}) = E(r)\mathbf{n}_r$ where $\mathbf{n}_r \equiv \mathbf{r}/r$ is the unit vector in the direction of the field observation point \mathbf{r} (Fig. 2).

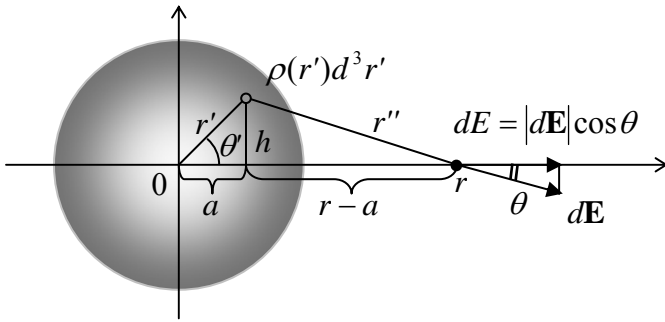


Fig. 1.2. One of the simplest problems of electrostatics: electric field produced by a spherically-symmetric charge distribution.

Taking this direction as the polar axis of a spherical coordinate system, we can use the evident independence of the elementary radial field dE , created by the elementary charge $\rho(r')d^3 r' = \rho(r')r'^2 \sin \theta' dr' d\theta' d\phi'$, of the azimuth angle ϕ' , and reduce integral (9) to

$$E = \frac{1}{4\pi\epsilon_0} 2\pi \int_0^\pi \sin \theta' d\theta' \int_0^\infty r'^2 dr' \frac{\rho(r')}{(r'')^2} \cos \theta, \quad (1.11)$$

¹⁰ Note that for a continuous, smooth charge distribution, integral (9) does not diverge at $\mathbf{R} \equiv \mathbf{r} - \mathbf{r}' \rightarrow 0$, because in this limit the fraction under the integral increases as R^{-2} , i.e. slower than the decrease of the elementary volume $d^3 r'$, proportional to R^3 .

¹¹ See, e.g., Sec. 14 of the *Selected Mathematical Formulas* appendix, referred below as MA.

where θ and r'' are the geometrical parameters marked in Fig. 2. Since they all may be readily expressed via r' and θ' using auxiliary parameters a and h ,

$$\cos \theta = \frac{r-a}{r''}, \quad (r'')^2 = h^2 + (r - r' \cos \theta)^2, \quad a = r' \cos \theta', \quad h = r' \sin \theta', \quad (1.12)$$

integral (11) may be eventually reduced to an explicit integral over r' and θ' . and worked out analytically, but that would require some effort.

For more complex problem, integral (8) may be much more complex, defying an analytical solution. One could argue that with the present-day abundance of computers and numerical algorithm libraries, one can always resort to numerical integration. This argument may be enhanced by the fact that numerical *integration* is based on the replacement of the integral by a sum, and summation is much more robust to (unavoidable) discretization and rounding errors than the finite-difference schemes typical for the numerical solution of *differential* equations.

These arguments, however, are only partly justified, since in many cases the numerical approach runs into a problem sometimes called the *curse of dimensionality*, in which the last word refers to the number of input parameters of the problem to be solved, i.e. the dimensionality of its parameter space. Let us discuss this issue, because it is common for most fields of physics and, more generally, any quantitative science.¹²

If the number of the parameters of a problem is small, the results of its numerical solution may be of the same (and in some sense higher) value than the analytical ones. For example, if a problem has no parameters, and its result is just one number (say, $\pi^2/4$), this “analytical” answer hardly carries more information than its numerical form 2.4674011... Now, if a problem has one input parameter (say, a), the result of an analytical approach in most cases may be presented as an analytical function $f(a)$. If the function is really simple, called elementary, with many properties well known (say, $f(a) = \sin a$), this function gives us virtually everything we want to know. However, if the function is complicated, you would need to tabulate it numerically for a set of values of parameter a and possibly present the result as a plot. The same results (and the same plot) can be calculated numerically, without using analytics at all. This plot may certainly be very valuable, but since the analytical form has a potential of giving you more information (say, the values of $f(a)$ outside the plot range, or the asymptotic behavior of the function), it is hard to say that the numerics completely beat the analytics here.

Now let us assume that you have more input parameters. For two parameters (say, a and b), instead of one curve you would need a family of such curves for several (sometimes many) values of b . Still, the plots sometimes may fit one page convenient for viewing, so it is still not too bad. Now, if you have three parameters, the full representation of the results may require many pages (maybe a book) full of curves, for four parameters we may speak about several bookshelves, for five parameters something like a library, etc. For large number of parameters, typical for many scientific problems, the number of points in the parameters space grows exponentially, even the volume of calculations necessary for the generation of this data may become impracticable, despite the dirt-cheap CPU time we have now.

Thus, despite the current proliferation of numerical methods in physics, analytical results have an ever-lasting value, and we should try to get them whenever we can. For our current problem of finding electric field generated by a fixed set of electric charges, large help comes from the *Gauss law*.

¹² Actually, the term “curse of dimensionality” was coined in the 1950s by R. Bellman in the context of the optimal control theory, and only later spread to other sciences that heavily rely on numerical calculations.

Let us consider a single point charge q inside a smooth, closed surface A (Fig. 3), and calculate product $E_n d^2r$, where d^2r is an infinitesimal element of the surface (which may be well approximated with a plane of that area), and E_n is the component of the electric field in that point, normal to that plane.

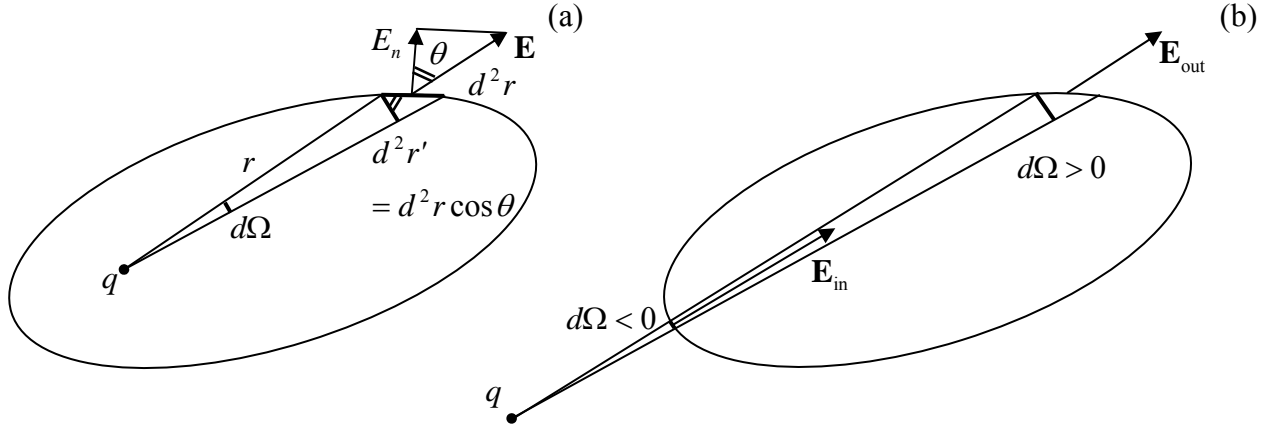


Fig. 1.3. Deriving the Gauss law: a point charge q is (a) inside volume V and (b) outside of that volume.

This component may be calculated as $E \cos \theta$, where θ is the angle between vector \mathbf{E} and the unit vector \mathbf{n} normal to the surface. (Equivalently, E_n may be presented as the scalar product $\mathbf{E} \cdot \mathbf{n}$.) Now let us notice that the product $\cos \theta d^2r$ is nothing more than the area d^2r' of the projection of d^2r onto the plane perpendicular to vector \mathbf{r} connecting charge q with this point of the surface (Fig. 3), because the angle between the planes d^2r' and d^2r is also equal to θ . Using the Coulomb law for \mathbf{E} , we get

$$E_n d^2r = E \cos \theta d^2r = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2} d^2r'. \quad (1.13)$$

But the ratio d^2r'/r^2 is nothing more than the elementary solid angle $d\Omega$ under which the areas d^2r' and d^2r are seen from the charge point, so that $E_n d^2r$ may be presented as just a product of $d\Omega$ by a constant ($q/4\pi\epsilon_0$). Summing these products over the whole surface, we get

$$\oint_S E_n d^2r = \frac{q}{4\pi\epsilon_0} \oint_S d\Omega = \frac{q}{\epsilon_0}, \quad (1.14)$$

since the full solid angle equals 4π . (The integral in the left-hand part of this relation is called the *flux of electric field* through surface S .)

Equation (14) expresses the Gauss law for one point charge. However, it is only valid if the charge is located *inside* the volume limited by the surface. In order to find the flux created by a charge *outside* of the volume, we still can use Eq. (13), but to proceed we have to be careful with the signs of the elementary contributions $E_n dA$. Let us use the common convention to direct the unit vector \mathbf{n} out of the closed volume we are considering (the so-called *outer normal*), so that the elementary product $E_n d^2r = (\mathbf{E} \cdot \mathbf{n}) d^2r$ and hence $d\Omega = E_n d^2r'/r^2$ is positive if vector \mathbf{E} is pointing out of the volume (like in the example shown in Fig. 3a and the upper-right area in Fig. 3b), and negative in the opposite case (for example, in the lower-left area in Fig. 3b). As the latter figure shows, if the charge is located outside of the volume, for each positive contribution $d\Omega$ there is always equal and opposite contribution to the

integral. As a result, at the integration over the solid angle the positive and negative contributions cancel exactly, so that

$$\oint_S E_n d^2r = 0. \quad (1.15)$$

The real power of the Gauss law is revealed by its generalization to the case of many charges within volume V . Since the calculation of flux is a linear operation, the linear superposition principle (3) means that the flux created by several charges is equal to the (algebraic) sum of individual fluxes from each charge, for which either Eq. (14) or Eq. (15) are valid, depending on the charge position (in or out of the volume). As the result, for the total flux we get:

Gauss
law

$$\oint_S E_n d^2r = \frac{Q_V}{\epsilon_0} \equiv \frac{1}{\epsilon_0} \sum_{\mathbf{r}_j \in V} q_j = \frac{1}{\epsilon_0} \int_V \rho(\mathbf{r}') d^3r', \quad (1.16)$$

where Q_V is the net charge inside volume V . This is the full version of the Gauss law.

In order to appreciate the problem-solving power of the law, let us return to the problem presented in Fig. 2, i.e. a spherical charge distribution. Due to its symmetry, which had already been discussed above, if we apply Eq. (16) to a sphere of radius r , the electric field should be perpendicular to the sphere at each its point (i.e., $E_n = E$), and its magnitude the same at all points: $E_n = E = E(r)$. As a result, the flux calculation is elementary:

$$\oint_S E_n d^2r = 4\pi r^2 E(r). \quad (1.17)$$

Now, applying the Gauss law (16), we get:

$$4\pi r^2 E(r) = \frac{1}{\epsilon_0} \int_{r' < r} \rho(r') d^3r' = \frac{4\pi}{\epsilon_0} \int_0^r r'^2 \rho(r') dr', \quad (1.18)$$

so that, finally,

$$E(r) = \frac{1}{r^2 \epsilon_0} \int_0^r r'^2 \rho(r') dr' = \frac{1}{4\pi \epsilon_0} \frac{Q(r)}{r^2}, \quad (1.19)$$

where $Q(r)$ is the full charge inside the sphere of radius r :

$$Q(r) \equiv 4\pi \int_0^r \rho(r') r'^2 dr'. \quad (1.20)$$

In particular, this formula shows that the field *outside* of a sphere of a finite radius R is exactly the same as if all its charge $Q = Q(R)$ was concentrated in the sphere's center. (Note that this important result is only valid for any spherically-symmetric charge distribution.) For the field *inside* the sphere, finding electric field still requires an explicit integration (20), but this 1D integral is much simpler than the 2D integral (11), and in some important cases may be readily worked out analytically. For example, if charge Q is uniformly distributed inside a sphere of radius R ,

$$\rho(r') = \rho = \frac{Q}{V} = \frac{Q}{(4\pi/3)R^3}, \quad (1.21)$$

the integration is elementary:

$$E(r) = \frac{\rho}{r^2 \varepsilon_0} \int_0^r r'^2 dr' = \frac{\rho r}{3 \varepsilon_0} = \frac{1}{4\pi \varepsilon_0} \frac{Qr}{R^3}. \quad (1.22)$$

We see that in this case the field is growing linearly from the center to the sphere's surface, and only at $r > R$ starts to decrease in agreement with Eq. (19) with constant $Q(r) = Q$. Another important observation is that the results for $r \leq R$ and $r \geq R$ give the same value ($Q/4\pi\varepsilon_0 R^2$) at the charged sphere's surface, $r = R$, so that the electric field is continuous.

In order to underline the importance of the last fact, let us consider one more elementary but very important example of the Gauss law's application. Let a thin plane sheet (Fig. 4) be charged uniformly, with an areal density $\sigma = \text{const}$ (see Footnote 9 above).

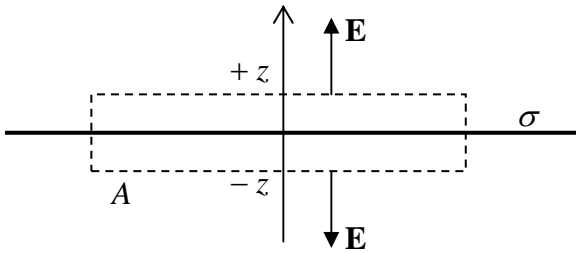


Fig. 1.4. Electric field of a charged plane.

In this case, it is fruitful to use the Gauss volume in the form of a planar “pillbox” of thickness $2z$ (where z is the Cartesian coordinate perpendicular to charged plane) and certain area A – see Fig. 4. Due to the symmetry of the problem, it is evident that the electric field should be: (i) directed along axis z , (ii) constant on each of the upper and bottom sides of the pillbox, (iii) equal and opposite on these sides, and (iv) parallel to the side surfaces of the box. As a result, the full electric field flux through the pillbox surface is just $2AE(z)$, so that the Gauss law (16) yields

$$2AE(z) = \frac{1}{\varepsilon_0} Q_A = \frac{1}{\varepsilon_0} \sigma A, \quad (1.23)$$

and we get a very simple but important formula

$$E(z) = \frac{\sigma}{2\varepsilon_0} = \text{const.} \quad (1.24)$$

Notice that, somewhat counter-intuitively, the field magnitude does not depend on the distance from the charged plane. From the point of view of the Coulomb law (5), this result may be explained as follows, the farther the observation point from the plane, the weaker the effect of each elementary charge, $dQ = \sigma d^2r$, but the more such elementary charges give contributions to the vertical component of vector \mathbf{E} .

Note also that though the magnitude $E \equiv |\mathbf{E}|$ of the electric field is constant, its vertical component E_z changes sign at $z = 0$ (Fig. 4), experiencing a *discontinuity* (jump) equal to $\Delta E_z = \sigma/\varepsilon_0$. This jump disappears if the surface is not charged ($\sigma = 0$). This statement remains true in a more general case of finite volume (but not surface!) charge density ρ . Returning for a minute to our charged

sphere problem, very close to its surface it may be considered planar, so that the electric field should indeed be continuous, as it is.

Admittedly, the *integral form* (16) of the Gauss law is immediately useful only for highly symmetrical geometries, like as in the two problems discussed above. However, it may be recast into an alternative, *differential form* whose field of useful applications is much wider. This form may be obtained from Eq. (16) using the *divergence theorem* that, according to the vector algebra, is valid for any space-differentiable vector, in particular \mathbf{E} , and for any volume V limited by closed surface S :¹³

$$\oint_S \mathbf{E} \cdot d\mathbf{r} = \int_V (\nabla \cdot \mathbf{E}) d^3r, \quad (1.25)$$

where ∇ is the *del* (or “nabla”) *operator* of spatial differentiation.¹⁴ Combining Eq. (25) with the Gauss law (16), we get

$$\int_V \left(\nabla \cdot \mathbf{E} - \frac{\rho}{\epsilon_0} \right) d^3r = 0. \quad (1.26)$$

For a given distribution of electric charge (and hence of the electric field), this equation should be valid for any choice of volume V . This can hold only if the function under the integral vanishes at each point, i.e. if¹⁵

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0}. \quad (1.27)$$

Note that in a sharp contrast with the integral form (16), Eq. (26) is *local*: it relates the electric field divergence to the charge density *at the same point*. This equation, being the differential form of the Gauss law, is frequently called (the free-space version of) one of *Maxwell equations*. Another, homogeneous Maxwell equation’s “embryo” may be obtained by noticing that curl of point charge’s field, and hence that of any system of charges, equals zero:¹⁶

$$\nabla \times \mathbf{E} = 0. \quad (1.28)$$

(We will arrive at two other Maxwell equations, for the magnetic field, in Chapter 5, and then generalize all the equations to their full, time-dependent form by the end of Chapter 6. However, Eq. (27) would stay the same.)

Just to get a better gut feeling of Eq. (27), let us apply it to the same example of a uniformly charged sphere (Fig. 2). The vector algebra tells us that the divergence of a spherically symmetric vector function $\mathbf{E}(\mathbf{r}) = E(r)\mathbf{n}_r$ may be simply expressed in spherical coordinates:¹⁷

¹³ See, e.g., MA Eq. (12.2). Note that the scalar product under the integral in Eq. (25) is nothing more than the divergence of vector \mathbf{E} – see, e.g., MA Eq. (8.4).

¹⁴ See, e.g., MA Secs. 8-10.

¹⁵ In the Gaussian units, just as in the initial Eq. (5), ϵ_0 has to be replaced with $1/4\pi$, so that the Maxwell equation (27) looks like $\nabla \cdot \mathbf{E} = 4\pi\rho$, while Eq. (28) stays the same.

¹⁶ This follows, for example, from the direct application of MA Eq. (10.11) to the spherically-symmetric vector function $\mathbf{f} = \mathbf{E}(\mathbf{r}) = E(r)\mathbf{n}_r$ field of a point charge placed at the origin, giving $f_\theta = f_\phi = 0$ and $\partial f_r / \partial \theta = \partial f_r / \partial \phi = 0$.

¹⁷ See, e.g., MA Eq. (10.10) for this particular case (when $\partial/\partial\theta = \partial/\partial\phi = 0$).

Inhomogeneous Maxwell equation for \mathbf{E}

Homogeneous Maxwell equation for \mathbf{E}

$$\nabla \cdot \mathbf{E} = \frac{1}{r^2} \frac{d}{dr} (r^2 E). \quad (1.29)$$

As a result, Eq. (27) yields a linear, ordinary differential equation for the function $E(r)$:

$$\frac{1}{r^2} \frac{d}{dr} (r^2 E) = \begin{cases} \rho / \varepsilon_0, & \text{for } r \leq R, \\ 0, & \text{for } r \geq R, \end{cases} \quad (1.30a)$$

that may be readily integrated on each of the segments:

$$E(r) = \frac{1}{\varepsilon_0} \frac{1}{r^2} \times \begin{cases} \rho \int r^2 dr = \rho r^3 / 3 + C_1, & \text{for } r \leq R, \\ C_2, & \text{for } r \geq R. \end{cases} \quad (1.30b)$$

In order to determine the integration constant C_1 , we can use boundary condition $E(0) = 0$. (It follows from problem's spherical symmetry: in the center of the sphere, electric field has to vanish, because otherwise, where would it be directed?) Constant C_2 may be found from the continuity condition $E(R - 0) = E(R + 0)$, which has already been discussed above. As a result, we arrive at our previous results (19) and (22).

We can see that in this particular, highly symmetric case, using the differential form of the Gauss law is more complex than its integral form. (For our second example, shown in Fig. 4, it would be even less natural.) However, Eq. (27) and its generalizations are more convenient for asymmetric charge distributions, and invaluable in the cases where the charge distribution $\rho(\mathbf{r})$ is not known a priori and has to be found in a self-consistent way. (We will start discussing such cases in the next chapter.)

1.3. Scalar potential and electric field energy

One more help for solving electrostatics (and more complex) problems may be obtained from the notion of the *electrostatic potential*, which is just the electrostatic potential energy U of a probe particle, normalized by its charge:

$$\phi \equiv \frac{U}{q}.$$

(1.31)

Electro-
static
potential

As we know from classical mechanics,¹⁸ the notion of U (and hence ϕ) make sense only for the case of *potential forces*, for example those depending just on particle's position. Equations (6) and (8) show that, in the static situations, the electric field clearly falls into this category. For such a field, the potential energy may be defined as a scalar function $U(\mathbf{r})$ that allows the force to be calculated as its gradient (with the opposite sign):

$$\mathbf{F} = -\nabla U. \quad (1.32)$$

Dividing both sides of this equation by the charge of the probe particle, and using Eqs. (5) and (31), we get¹⁹

¹⁸ See, e.g., CM Sec. 1.4.

¹⁹ Eq. (28) could be also derived from this relation, because according to vector algebra, any gradient field has vanishing curl - see, e.g., MA Eq. (11.1).

Electrostatic
field as a
gradient

$$\mathbf{E} = -\nabla\phi. \quad (1.33)$$

In order to calculate the scalar potential, let us start from the simplest case of a single point charge q placed at the origin. For it, the Coulomb law (5) takes a simple form

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0} q \frac{\mathbf{r}}{r^3} = \frac{1}{4\pi\epsilon_0} q \frac{\mathbf{n}_r}{r^2}. \quad (1.34)$$

It is straightforward to check that the last fraction in the right-hand part of this equation is equal to $-\nabla(1/r)$.²⁰ Hence, according to the definition (33), for this particular case

$$\phi = \frac{1}{4\pi\epsilon_0} \frac{q}{r}. \quad (1.35)$$

Potential of a
point charge

(In the Gaussian units, this result is spectacularly simple: $\phi = q/r$.) Note that we could add an arbitrary constant to this potential (and indeed to *any* other distribution of ϕ discussed below) without changing the force, but it is convenient to define the potential energy to approach zero at infinity.

Before going any further, let us demonstrate how useful the notions of U and ϕ are, on a very simple example. Let two similar charges q be launched from afar, with an initial velocity $v_0 \ll c$ each, straight toward each other (i.e. with the zero impact parameter) – see Fig. 5. Since, according to the Coulomb law, the charges repel each other with increasing force, they will stop at some minimum distance r_{\min} from each other, and then fly back.

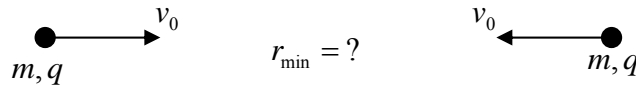


Fig. 1.5. Simple problem of electric particle motion.

We could of course find r_{\min} directly from the Coulomb law. However, for that we would need to write the 2nd Newton law for each particle (actually, due to the problem symmetry, they would be similar), then integrate them over time once to find the particle velocity v as a function of distance, and then recover r_{\min} from the requirement $v = 0$. The notion of potential allows this problem to be solved in one line. Indeed, in the field of potential forces the system's total energy $E = T + U = T + q\phi$ is conserved. In our nonrelativistic case, the kinetic energy T is just $mv^2/2$. Hence, equating the total energy of two particles in the points $r = \infty$ and $r = r_{\min}$, and using Eq. (35) for ϕ , we get

$$2 \frac{mv_0^2}{2} + 0 = 0 + \frac{1}{4\pi\epsilon_0} \frac{q^2}{r_{\min}}, \quad (1.36)$$

immediately giving us the final answer: $r_{\min} = q^2/4\pi\epsilon_0 mv_0^2$.

Now let us calculate ϕ for an arbitrary configuration of charges. For a single charge in an arbitrary position (say, \mathbf{r}_k), r in Eq. (35) should be evidently replaced for $|\mathbf{r} - \mathbf{r}_k|$. Now, the linear

²⁰ This may be done either by Cartesian components or using the well-known expression $\nabla f = (df/dr)\mathbf{n}_r$, valid for any spherically-symmetric scalar function $f(r)$ - see, e.g., MA Eq. (10.8) for the particular case $\partial/\partial\theta = \partial/\partial\varphi = 0$.

superposition principle (3) allows for an easy generalization of this formula to the case of an arbitrary set of discrete charges,

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \sum_{\mathbf{r}_{k'} \neq \mathbf{r}} \frac{q_{k'}}{|\mathbf{r} - \mathbf{r}_{k'}|}. \quad (1.37)$$

Finally, using the same arguments as in Sec. 1, we can use this result to argue that in the case of an arbitrary continuous charge distribution

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3 r'. \quad (1.38)$$

Potential
of a charge
distribution

Again, the notion of Dirac's delta-function allows to use the last equation for discrete charges as well, so that Eq. (38) may be considered as the general expression for the electrostatic potential.

For most practical calculations, using this expression and then applying Eq. (33) to the result, is preferable to using Eq. (9), because ϕ is a scalar, while \mathbf{E} is a 3D vector - mathematically equivalent to 3 scalars. Still, this approach may lead to technical problems similar to those discussed in Sec. 2. For example, applying it to the spherically-symmetric distribution of charge (Fig. 2), we get integral

$$\phi = \frac{1}{4\pi\epsilon_0} 2\pi \int_0^\pi \sin \theta' d\theta' \int_0^\infty r'^2 dr' \frac{\rho(r')}{r''} \cos \theta, \quad (1.39)$$

which is not much simpler than Eq. (11).

The situation may be much improved by re-casting Eq. (38) into a differential form. For that, it is sufficient to plug the definition of ϕ , Eq. (33), into Eq. (27):

$$\nabla \cdot (-\nabla \phi) = \frac{\rho}{\epsilon_0}. \quad (1.40)$$

The left-hand part of this equation is nothing more than the Laplace operator of ϕ (with the minus sign), so that we get the famous *Poisson equation*²¹ for the electrostatic potential:

$$\nabla^2 \phi = -\frac{\rho}{\epsilon_0}. \quad (1.41)$$

Poisson
equation
for ϕ

(In the Gaussian units, the Poisson equation looks like $\nabla^2 \phi = -4\pi\rho$.) This differential equation is so convenient for applications that even its particular case for $\rho = 0$,

$$\nabla^2 \phi = 0, \quad (1.42)$$

Laplace
equation
for ϕ

has earned a special name – the *Laplace equation*.²²

In order to get a feeling of the Poisson equation as a problem solving tool, let us return to the spherically-symmetric charge distribution (Fig. 2) with a constant charge density ρ . Using the

²¹ Named after S. D. Poisson (1781-1840), also famous for the *Poisson distribution* – one of the central results of the probability theory - see, e.g., SM Sec. 5.2.

²² After mathematician (and astronomer) P. S. de Laplace (1749-1827) who, together with A. Clairault, is credited for the development of the very concept of potential.

symmetry, we can present the potential as $\phi(\mathbf{r}) = \phi(r)$, and hence use the following simple expression for its Laplace operator:²³

$$\nabla^2 \phi = \frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{d\phi}{dr} \right), \quad (1.43)$$

so that for the points inside the charged sphere ($r \leq R$) the Poisson equation yields

$$\frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{d\phi}{dr} \right) = -\frac{\rho}{\varepsilon_0}, \quad \text{i.e.} \quad \frac{d}{dr} \left(r^2 \frac{d\phi}{dr} \right) = -\frac{\rho}{\varepsilon_0} r^2. \quad (1.44)$$

Integrating the last form of the equation over r once, with the natural boundary condition $d\phi/dr|_{r=0} = 0$ (because of the condition $E(0) = 0$, which has been discussed above), we get

$$\frac{d\phi}{dr}(r) = -\frac{\rho}{r^2 \varepsilon_0} \int_0^r r'^2 dr' = -\frac{\rho r}{3\varepsilon_0} = -\frac{1}{4\pi\varepsilon_0} \frac{Qr}{R^3}. \quad (1.45)$$

Since this derivative is nothing more than $-E(r)$, in this formula we can readily recognize our previous result (22). Now we may like to carry out the second integration to calculate the potential itself:

$$\phi(r) = -\frac{Q}{4\pi\varepsilon_0 R^3} \int_0^r r' dr' + c_1 = -\frac{Qr^2}{8\pi\varepsilon_0 R^3} + c_1. \quad (1.46)$$

Before making any judgment on the integration constant c_1 , let us solve the Poisson equation (in this case, just the Laplace equation) for the range outside the sphere ($r > R$):

$$\frac{1}{r^2} \frac{d}{dr} \left(r^2 \frac{d\phi}{dr} \right) = 0. \quad (1.47)$$

Its first integral,

$$\frac{d\phi}{dr}(r) = \frac{c_2}{r^2}, \quad (1.48)$$

also gives the electric field (with the minus sign). Now using Eq. (1.45) and requiring the field to be continuous at $r = R$, we get

$$\frac{c_2}{R^2} = -\frac{Q}{4\pi\varepsilon_0 R^2}, \quad \text{i.e.} \quad \frac{d\phi}{dr}(r) = -\frac{Q}{4\pi\varepsilon_0 r^2}, \quad (1.49)$$

in an evident agreement with Eq. (19). Integrating this result again,

$$\phi(r) = -\frac{Q}{4\pi\varepsilon_0} \int \frac{dr}{r^2} = \frac{Q}{4\pi\varepsilon_0 r} + c_3, \quad \text{for } r > R, \quad (1.50)$$

we can select $c_3 = 0$, so that $\phi(\infty) = 0$, in accordance with the usual (though not compulsory) convention. Now we can finally determine constant c_1 in Eq. (46) by requiring that this equation and Eq. (50) give the same value of ϕ at the boundary $r = R$. (According to Eq. (33), if the potential had a jump, the electric field at that point would be infinite.) The final answer may be presented as

²³ See, e.g., MA Eq. (10.8) for $\partial/\partial\theta = \partial/\partial\varphi = 0$.

$$\phi(r) = \frac{Q}{4\pi\epsilon_0 R} \left[\frac{R^2 - r^2}{2R^2} + 1 \right], \quad \text{for } r \leq R. \quad (1.51)$$

We see that using the Poisson equation to find the electrostatic potential distribution for highly symmetric problems may be more cumbersome than directly finding the electric field – say, from the Gauss law. However, we will repeatedly see below that if the electric charge distribution is not fixed in advance, using Eq. (41) may be the only practicable way to proceed.

Returning now to the general theory of electrostatic phenomena, let us calculate potential energy U of an arbitrary system of electric charges q_k . Despite the apparently straightforward relation (31) between U and ϕ , the calculation is a little bit more complex than one might think. Indeed, let us rewrite Eqs. (32), (33) for a *single* charge in the integral form:

$$U(\mathbf{r}) = - \int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{F}(\mathbf{r}') \cdot d\mathbf{r}', \quad \text{i.e.} \quad \phi(\mathbf{r}) = - \int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{E}(\mathbf{r}') \cdot d\mathbf{r}', \quad (1.52)$$

where \mathbf{r}_0 is some reference point. These integrals reflect the fact that the potential energy is just the work necessary to move the charge from point \mathbf{r}_0 to point \mathbf{r} , and clearly depend on whether the charge motion affects force \mathbf{F} (and hence electric field \mathbf{E}) or not. If it does not, i.e. if the field is produced by some *external* charges (such fields \mathbf{E}_{ext} are also called *external*), everything is simple indeed: using the linearity of relations (31) and (32), for the total potential energy we may write

$$U_{\text{ext}} = \sum_k q_k \phi_{\text{ext}}(\mathbf{r}_k), \quad \text{where } \phi_{\text{ext}}(\mathbf{r}) \equiv - \int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{E}_{\text{ext}}(\mathbf{r}') \cdot d\mathbf{r}'. \quad (1.53)$$

Repeating the argumentation that has led us to Eq. (9), we see that for a continuously distributed charge, this sum turns into an integral:

$$U_{\text{ext}} = \int \rho(\mathbf{r}) \phi_{\text{ext}}(\mathbf{r}) d^3r. \quad (1.54)$$

Energy
in
external
field

However, if the electric field is created by the charges whose energy we are calculating, the situation is somewhat different. To calculate U for this case, let us use the fact its independence of the way the charge configuration has been created, considering the following process. First, let us move one charged particle (say, q_1) from infinity to an arbitrary point of space (\mathbf{r}_1) in the absence of other charges. During the motion the particle does not experience any force (again, the charge does not interact with itself!), so that its potential energy is the same as at infinity (with the standard choice of the arbitrary constant, zero): $U_1 = 0$. Now let us fix the position of that charge, and move another charge (q_2) from infinity to point \mathbf{r}_2 (with velocity $v \ll c$, in order to avoid any magnetic field effects, to be discussed in Chapter 5.) This particle, during its motion, does experience the Coulomb force exerted by fixed q_1 , so that according to Eq. (31), its contribution to the final potential energy

$$U_2 = q_2 \phi_1(\mathbf{r}_2). \quad (1.55)$$

Since the first particle was not moving during this process, the total potential energy U of the system is equal to just U_2 . This is exactly the equality used for writing the right-hand part of Eq. (36). (Prescribing a similar energy to charge q_1 as well would constitute an error – a very popular one, and hence having a special name, *double-counting*.)

Now, fixing the first two charges in points \mathbf{r}_1 and \mathbf{r}_2 , respectively, and bringing in the third charge from infinity, we *increment* the potential energy by

$$U_3 = q_3[\phi_1(\mathbf{r}_3) + \phi_2(\mathbf{r}_3)]. \quad (1.56)$$

I believe that at this stage it is already clear how to generalize this result to the contribution from an arbitrary (k -th) charge being moved in (Fig. 6):

$$U_k = q_k[\phi_1(\mathbf{r}_k) + \phi_2(\mathbf{r}_k) + \phi_3(\mathbf{r}_k) + \dots + \phi_{k-1}(\mathbf{r}_k)] = q_k \sum_{k' < k} \phi_{k'}(\mathbf{r}_k). \quad (1.57)$$

(Notice condition $k' < k$, which suppresses erroneous double-counting.)

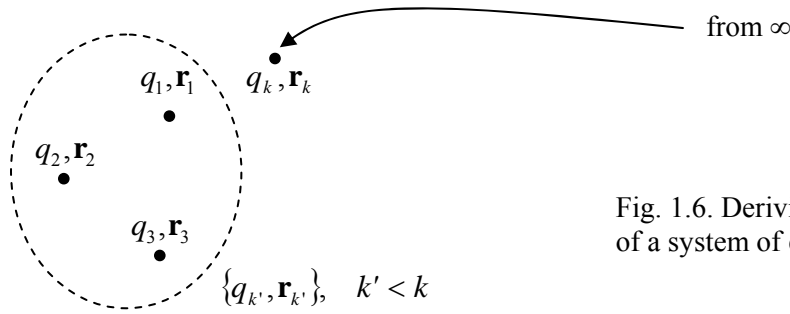


Fig. 1.6. Deriving the potential energy of a system of electric charges.

Now, summing up all the increments, for the total electrostatic energy of the system we get:

$$U = \sum_k U_k = \sum_{\substack{k, k' \\ (k' < k)}} q_k \phi_{k'}(\mathbf{r}_k). \quad (1.58)$$

This is our final result in its generic form; it is so important that it is worthy of rewriting it in two other forms. First, for its generalization to the continuous charge distribution, we may use Eq. (35) to present Eq. (58) in a more symmetric form:

$$U = \frac{1}{4\pi\epsilon_0} \sum_{\substack{k, k' \\ (k' < k)}} \frac{q_k q_{k'}}{|\mathbf{r}_k - \mathbf{r}_{k'}|}. \quad (1.59)$$

The expression under the sum is evidently symmetric with respect to the index swap, so that it may be rewritten in a fully symmetric form,

$$U = \frac{1}{4\pi\epsilon_0} \frac{1}{2} \sum_{\substack{k', k \\ (k' \neq k)}} \frac{q_k q_{k'}}{|\mathbf{r}_k - \mathbf{r}_{k'}|}, \quad (1.60)$$

which is now easily generalized to the continuous case:

$$U = \frac{1}{4\pi\epsilon_0} \frac{1}{2} \int d^3r \int d^3r' \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \quad (1.61)$$

(As before, in this case the restriction expressed in the discrete charge case as $k \neq k'$ is not important, because if the charge density is a continuous function, integral (61) does not diverge at point $\mathbf{r} = \mathbf{r}'$.)

To present this result in one more form, let us notice that according to Eq. (38), the integral over r' in Eq. (61), divided by $4\pi\epsilon_0$, is just the full electrostatic potential at point \mathbf{r} , and hence

$$U = \frac{1}{2} \int \rho(\mathbf{r}) \phi(\mathbf{r}) d^3r. \quad (1.62)$$

Charge
interaction
energy

For the discrete charge case, this result becomes

$$U = \frac{1}{2} \sum_k q_k \phi(\mathbf{r}_k), \quad (1.63)$$

but now it is important to remember that the “full” potential’s value $\phi(\mathbf{r}_k)$ should exclude the (infinite) contribution of charge k itself. Comparing the last two formulas with Eqs. (52) and (53), we see that the electrostatic energy of charge interaction, as expressed via the charge-potential product, is twice less than that of charge energy in a fixed (“external”) field. This is evidently the result of the self-consistent build-up of the electric field as the charge system is being formed.²⁴

Now comes an important conceptual question: can we locate this interaction energy in space? Expressions (60)-(63) seem to imply that contributions to U come only from the regions where electric charges are located. However, one of the beautiful features of physics is that sometimes completely different interpretations of the same mathematical result are possible. In order to get an alternative view at our current result, let us write Eq. (62) for a volume V so large that the electric field on the limiting surface A is negligible, and plug into it the charge density expressed from the Poisson equation (41):

$$U = -\frac{\epsilon_0}{2} \int_V \phi \nabla^2 \phi d^3r. \quad (1.64)$$

This expression may be integrated by parts as²⁵

$$U = -\frac{\epsilon_0}{2} \left[\oint_A \phi (\nabla \phi)_n d^2r - \int_V (\nabla \phi)^2 d^3r \right]. \quad (1.65)$$

According to our condition of negligible field $\mathbf{E} = -\nabla \phi$ on the surface, the first integral vanishes, and we get a very important formula

$$U = \frac{\epsilon_0}{2} \int (\nabla \phi)^2 d^3r = \frac{\epsilon_0}{2} \int E^2 d^3r. \quad (1.66)$$

This result certainly invites an interpretation very much different than Eq. (62): it is natural to represent it in the following form:

$$U = \int u(\mathbf{r}) d^3r, \quad \text{with } u(\mathbf{r}) \equiv \frac{\epsilon_0}{2} E^2(\mathbf{r}), \quad (1.67)$$

Electric
field
energy

²⁴ The nature of this additional factor $\frac{1}{2}$ is absolutely the same as in the well-known formula $U = (\frac{1}{2})\kappa x^2$ for the potential energy of an elastic spring providing returning force $F = -\kappa x$ proportional to the deviation x from equilibrium.

²⁵ This transformation follows from the divergence theorem MA (12.2) applied to vector function $\mathbf{f} = \phi \nabla \phi$, taking into account the 3D differentiation rule MA Eq. (11.4a): $\nabla \cdot (\phi \nabla \phi) = (\nabla \phi) \cdot (\nabla \phi) + \phi \nabla \cdot (\nabla \phi) = (\nabla \phi)^2 + \phi \nabla^2 \phi$.

and interpret $u(\mathbf{r})$ as the *spatial density of the electric field energy*,²⁶ which is continuously distributed over all the space where the field exists - rather than just its part where the charges are located.

Let us have a look how these two alternative pictures work for our testbed problem, a uniformly charged sphere. If we start from Eq. (62), we may limit integration by the sphere volume ($0 \leq r \leq R$) where $\rho \neq 0$. Using Eq. (51), and the spherical symmetry of the problem ($d^3r = 4\pi r^2 dr$), we get

$$U = \frac{1}{2} 4\pi \int_0^R \rho \phi r^2 dr = \frac{1}{2} 4\pi \rho \frac{Q}{4\pi\epsilon_0 R} \int_0^R \left[\frac{R^2 - r^2}{2R^2} + 1 \right] r^2 dr = \frac{6}{5} \frac{1}{4\pi\epsilon_0 R} \frac{Q^2}{2}. \quad (1.68)$$

On the other hand, if we use Eq. (67), we need to integrate energy everywhere, i.e. both inside and outside of the sphere:

$$U = \frac{\epsilon_0}{2} 4\pi \left[\int_0^R E^2 r^2 dr + \int_R^\infty E^2 r^2 dr \right]. \quad (1.69)$$

Using Eqs. (19) and (22) for, respectively, the external and internal regions, we get

$$U = \frac{\epsilon_0}{2} 4\pi \left[\int_0^R \left(\frac{Qr}{4\pi\epsilon_0} \right)^2 r^2 dr + \int_R^\infty \left(\frac{Q}{4\pi\epsilon_0 r^2} \right)^2 r^2 dr \right] = \left(\frac{1}{5} + 1 \right) \frac{1}{4\pi\epsilon_0 R} \frac{Q^2}{2}. \quad (1.70)$$

This is (fortunately :-)) the same answer as given by Eq. (68), but to some extent it is more informative because it shows how exactly the electric field energy is distributed between the interior and exterior of the charged sphere.²⁷

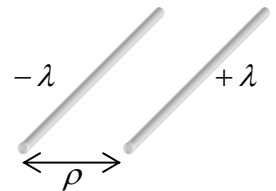
We see that, as we could expect, within the realm of *electrostatics*, Eqs. (62) and (67) are equivalent. However, when we examine *electrodynamics* in Chapter 6 and on, we will see that the latter equation is more general, and that it is more adequate to associate energy with the field itself rather than its sources - in our current case, electric charges.

1.4. Exercise problems

1.1. Calculate the electric field created by a thin, long, straight filament, electrically charged with a constant linear density λ , using two approaches:

- (i) directly from the Coulomb law, and
- (ii) using the Gauss law.

1.2. Two thin, straight parallel filaments, separated by distance ρ , carry equal and opposite uniformly distributed charges with linear density λ - see Fig. on the right. Calculate the electrostatic force (per unit length) of the Coulomb



²⁶ In the Gaussian units, the standard replacement $\epsilon_0 \rightarrow 1/4\pi$ turns the last of Eqs. (67) into $u(\mathbf{r}) = E^2/8\pi$.

²⁷ Note that $U \rightarrow \infty$ at $R \rightarrow 0$. Such divergence appears at application of Eq. (67) to any point charge. Since it does not affect the force acting on the charge, the divergence does not create any technical difficulty for analysis of charge statics or nonrelativistic dynamics, but it points to a conceptual problem of classical electrodynamics as the whole. This issue will be discussed in the very end of the course (Sec. 10.6).

interaction between the wires. Compare the result with the Coulomb law for the force between the point charges, and interpret their difference.

1.3. A sphere of radius R , whose volume had been charged with a constant density ρ , is split with a very narrow, planar gap passing through its center. Find the Coulomb force between the resulting two hemispheres.

1.4. Calculate the distribution of the electrostatic potential created by a straight, thin filament of finite length $2l$, charged with a constant linear density λ , and explore the result in the limits of very small and very large distances from the filament.

1.5. A thin plane sheet, perhaps of an irregular shape, carries an electric charge distributed over the sheet with a constant areal density σ .

(i) Express the electric field component normal to the plane, at a certain distance from it, via the solid angle Ω at which the sheet is visible from the observation point.

(ii) Use the result to calculate the field in the center of a cube, with one face charged with constant density σ .

1.6. Can one create electrostatic fields with the Cartesian components proportional to the following products of Cartesian coordinates $\{x, y, z\}$,

$$(i) \quad \{yz, xz, xy\},$$

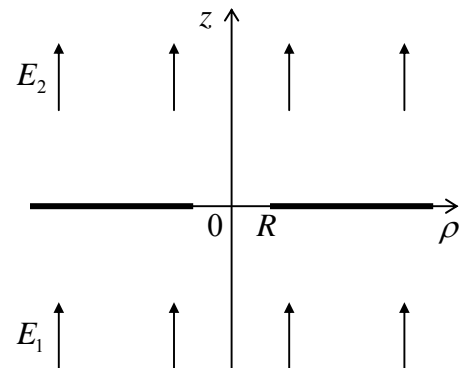
$$(ii) \quad \{xy, xy, yz\},$$

in a finite region of space?

1.7. Distant sources have been used to create different electric fields on two sides of a wide and thin metallic membrane with a round hole of radius R in it - see Fig. on the right. Besides the local perturbation created by the hole, the fields are uniform:

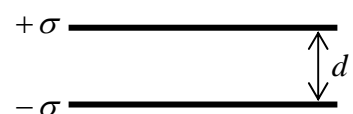
$$E|_{r \gg R} = \mathbf{n}_z \times \begin{cases} E_1, & \text{at } z < 0, \\ E_2, & \text{at } z > 0. \end{cases}$$

Prove that the system may serve as an electrostatic lens for charged particles flying along axis z , at distances $\rho \ll R$ from it, and calculate the focal distance f of the lens. Spell out the conditions of validity of your result.



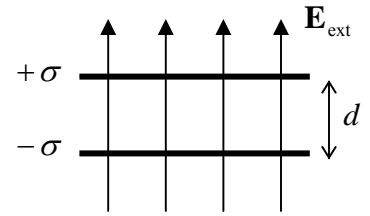
1.8. By direct calculation, find the average electric potential of the spherical surface of radius R , created by a point charge q located at distance $r > R$ from the sphere's center. Use the result to prove the following general *mean value theorem*: the electric potential at any point is always equal to its average value on any spherical surface with the center at that point, and containing no electric charges inside it.

1.9. Calculate the electrostatic energy per unit area of the system of two thin, parallel planes with equal and opposite charges of a constant areal density σ , separated by distance d - see Fig. on the right.



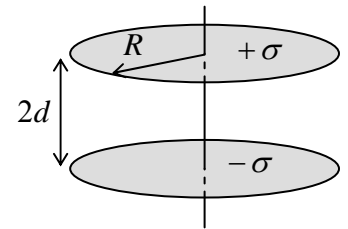
1.10. The system analyzed in the previous problem (two thin, parallel, oppositely charged planes) is now placed into an external, uniform, normal electric field $E_{\text{ext}} = \sigma/\epsilon_0$ – see Fig. on the right. Find the forces (per unit area) acting on each plane, by two methods:

- (i) directly from the electric field distribution, and
- (ii) from the potential energy of the system.



1.11. A thin spherical shell of radius R , which had been charged with a constant areal density σ , is split into two equal halves by a very narrow, planar cut passing through sphere's center. Calculate the force of electrostatic repulsion between the resulting hemispheric shells.

1.12. Two similar thin, circular, coaxial disks of radius R , separated by distance $2d$, are uniformly charged with equal and opposite areal densities $\pm\sigma$ – see Fig. on the right. Calculate and sketch the distribution of the electrostatic potential and the electric field of the disks along their common axis.



1.13. In a certain reference frame, the electrostatic potential created by some electric charge distribution, is

$$\phi(\mathbf{r}) = C \left(\frac{1}{r} + \frac{1}{2r_0} \right) \exp \left\{ -\frac{r}{r_0} \right\},$$

where C and r_0 are constants, and $r \equiv |\mathbf{r}|$ is the distance from the origin. Calculate the charge distribution in space.

1.14. A thin flat sheet, cut in a form of a rectangle of size $a \times b$, is electrically charged with a constant areal density σ . Without an explicit calculation of the spatial distribution $\phi(\mathbf{r})$ of the electrostatic potential induced by this charge, find the ratio of its values at the center and at the corners of the rectangle.

Hint: Consider partitioning the rectangle into several similar parts and using the linear superposition principle.

1.15. Explore the relation between the Laplace equation (42) and the condition of minimum of the electrostatic field energy (67).

1.16. Calculate the energy of electrostatic interaction of two spheres, of radii R_1 and R_2 , each with a spherically-symmetric charge distribution, separated by distance $d > R_1 + R_2$.

1.17. Prove the following *reciprocity theorem of electrostatics*:²⁸ if two spatially-confined charge distributions $\rho_1(\mathbf{r})$ and $\rho_2(\mathbf{r})$ create respective distributions $\phi_1(\mathbf{r})$ and $\phi_2(\mathbf{r})$ of the electrostatic potential, then

$$\int \rho_1(\mathbf{r})\phi_2(\mathbf{r})d^3r = \int \rho_2(\mathbf{r})\phi_1(\mathbf{r})d^3r.$$

Hint: Consider integral $\int \mathbf{E}_1 \cdot \mathbf{E}_2 d^3r$.

²⁸ This is only the simplest one of the whole family of reciprocity theorems in electromagnetism. (Sometimes it is called "Green's reciprocity theorem", but historically it is more fair to reserve the last name for the generalization to surface charges, using Eq. (2.210), to be discussed in Sec. 2. 7 below.)

Chapter 2. Charges and Conductors

In this chapter I will start addressing the (very common) situations when the electric charge distribution in space is not known a priori, but rather should be calculated in a self-consistent way together with the electric field it creates. The simplest situations of this kind involve conductors, and lead to the so-called boundary problems in which partial differential equations are solved with appropriate boundary conditions. Such problems are also broadly used in other parts of electrodynamics (and indeed in other fields of physics as well), so that following tradition, I will use this chapter's material as a playground for a discussion of various methods of boundary problem solution, and the special functions most frequently encountered on this way.

2.1. Electric field screening

The basic principles of electrostatics outlined in Chapter 1 present the conceptually full solution for the problem of finding electric field (and hence Coulomb forces) induced by a charge distribution, for example, charge density $\rho(\mathbf{r})$. However, in most practical situation this function is not known but should be found self-consistently with the field. The conceptually simplest case of this type arises when certain point charges q_k are placed near a surface of a good conductor, e.g., a metal: the electric field of these charges induces additional charges at conductor's surface, which also contribute to the field. Another important type of problems are those without space-positioned charges at all; here only the total charges of the involved conductors are fixed, but their spatial distribution inside each conductor has to be found. The full solution of such problems, of course, should satisfy Eq. (1.5) for the total field and total set of charges.

To approach the problems, I need to discuss, if only very briefly,¹ the relevant physics of conductors. In the simplest *macroscopic model*, conductors are treated as materials having internal charged particles (e.g., electrons in metals) that are free to move under the effect of force – in particular, the force $\mathbf{F} = q\mathbf{E}$ exerted by electric field \mathbf{E} . In electrostatics (which specifically excludes the case dc current, to be discussed in Chapter 4 below), there should be no such motion, so that everywhere inside the conductor the electric field should vanish:

$$\mathbf{E} = 0. \quad (2.1a)$$

Conductor's
interior in
electrostatics

This is the *electric field screening*² effect. According to Eq. (1.33), this condition may be rewritten in another, frequently more convenient form:

$$\phi = \text{const} ; \quad (2.1b)$$

note, however, that if a problem includes several unconnected conductors, the constant in Eq. (1b) may be different for each of them.

¹ More detailed discussions may be found, e.g., in Sec. 13.5 of J. Hook and H. Hall, *Solid State Physics*, 2nd ed., Wiley, 1991, or the section on electric field screening in Chapter 17 of N. Ashcroft and N. Mermin, *Solid State Physics*, Brooks Cole, 1976.

² This term, used for *electric* field, should not be confused with *shielding* – the word used for the description of *magnetic* field reduction by magnetic materials – see Chapter 5 below.

Now let us examine what we can say about the electric field *outside* a conductor, within the same macroscopic model. At close proximity, any smooth surface (in our case that of a conductor) looks planar. Let us integrate Eq. (1.28) over a narrow ($d \ll l$) rectangular loop C encircling a part of such plane conductor's surface (see the dashed line in Fig. 1), and apply to the electric field the well-known vector algebra equality - the *Stokes theorem*³

$$\oint_S (\nabla \times \mathbf{E})_n d^2r = \oint_C \mathbf{E} \cdot d\mathbf{r}, \quad (2.2)$$

where S is the surface limited by contour C , in our case dominated by two straight lines of length l . This means that if l is much smaller than the characteristic scale of field change, the right-hand part of Eq. (2) equals $[(E_\tau)_{\text{in}} - (E_\tau)_{\text{out}}]l$, where E_τ is field's component parallel to the surface. On the other hand, according to Eq. (1.28), the left-hand side of Eq. (2) equals zero. Hence, E_τ should be continuous at the surface, and in order to satisfy Eq. (1a) inside the conductor, immediately outside it, $E_\tau = 0$ as well.

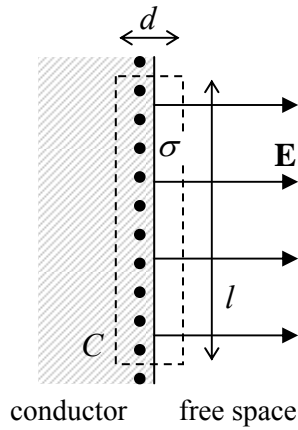


Fig. 2.1. Electric field near conductor's surface:
 $E_\tau = 0$, $E_n = \sigma/\epsilon_0$.

Hence, the field just outside the conductor has to be normal to its surface. In order to find this normal field, let us apply the Gauss law (1.16) to a plane pillbox of area A , similar to the one discussed in Sec. 1.2 – see Fig. 1.4. Due to Eq. (1), the total electric flux through the pillbox walls is now $(E_n)_{\text{out}}A$, so that for this surface field we get

$$\sigma = \epsilon_0 (E_n)_{\text{out}} = -\epsilon_0 (\nabla \phi)_n \equiv -\epsilon_0 \frac{\partial \phi}{\partial n}, \quad (2.3)$$

Surface
charge
density

where σ is the areal density of conductor's surface charge. So, the normal component of the field is related to the surface charge density by the universal relation (3).

For the electrostatic potential the macroscopic model provides an even more simple result. Indeed, applying the latter of integrals (1.52) to a short path d across the surface normal to it, we see that since E_n is finite, the potential change $\Delta\phi$ vanishes as $d \rightarrow 0$. Hence Eq. (1b) is also valid for potential's value immediately outside conductor's surface.

Before starting to use the macroscopic model for solution of particular problems of electrostatics, let us briefly discuss its limitations. Since the argumentation leading to Eq. (3) is valid for any thickness d of the Gauss pillbox, within the macroscopic model, the surface charge is located within an infinitely

³ See, e.g., MA Eq. (12.1).

thin surface layer. This is of course impossible physically: for one, this would require an infinite volume density ρ of charge. In reality the charged layer (and hence the region of electric field's crossover from the finite value (3) to zero) has a nonvanishing thickness λ . At least three effects contribute to λ :

(i) *Atomic structure of matter.* Within each atom, the electric field does exist and is highly non-uniform. Thus Eq. (1) is valid only for the *spatial average* of the field in a conductor, and cannot be taken seriously on the atomic scale $a_0 \sim 10^{-10}$ m.⁴

(ii) *Thermal excitation.* In conductor's bulk, the number of protons of atomic nuclei (n) and electrons (n_e) and per unit volume are balanced, so that the net charge density, $\rho = e(n - n_e)$, vanishes.⁵ However, if an external electric field penetrates a conductor, electrons can shift in or out of its affected part, depending on the field addition to their potential energy, $\Delta U = q_e \phi = -e\phi$. (Here the arbitrary constant in ϕ is chosen to give $\phi = 0$ inside the conductor.) In classical statistics, this change is described by the Boltzmann distribution:⁶

$$n_e(\mathbf{r}) = n \exp\left\{-\frac{U(\mathbf{r})}{k_B T}\right\}, \quad (2.4)$$

where $k_B \approx 1.38 \times 10^{-23}$ J/K is the Boltzmann constant, and T is temperature in SI units (kelvins). As a result, the net charge density is

$$\rho(\mathbf{r}) = en \left(1 - \exp\left\{\frac{e\phi(\mathbf{r})}{k_B T}\right\}\right). \quad (2.5)$$

If the field did not move the atomic nuclei at all, we could plug the last formula directly into the Poisson equation (1.49). Actually, the penetrating electric field shifts the average charge of the nuclei as well. As will be discussed in the next chapter, this results in the reduction of the electric field by a media-specific dimensionless factor ϵ_r (typically not too different from 1), called the *dielectric constant*. As a result, the Poisson equation takes the form,⁷

$$\frac{d^2 \phi}{dz^2} = -\frac{\rho}{\epsilon_r \epsilon_0} = \frac{en}{\epsilon_r \epsilon_0} \left(\exp\left\{\frac{e\phi}{k_B T}\right\} - 1\right), \quad (2.6)$$

where we have taken advantage of the 1D geometry of the system to simplify the Laplace operator, with axis z normal to the surface. Even with this simplification, Eq. (6) is a nonlinear differential equation allowing an analytical but rather bulky solution. Since our current goal is just to estimate of the field penetration depth λ , let us simplify the equation further by considering the low-field limit: $e|\phi| \sim e|E|\lambda \ll k_B T$. In this limit we can extend the exponent into the Taylor series, and limit ourselves to the two leading terms (of which the first one cancels with the unity). As a result, Eq. (6) becomes linear,

$$\frac{d^2 \phi}{dz^2} = \frac{en}{\epsilon \epsilon_0} \frac{e\phi}{k_B T}, \quad \text{i.e.} \quad \frac{d^2 \phi}{dz^2} = \frac{1}{\lambda^2} \phi, \quad (2.7)$$

⁴ This scale originates from the quantum-mechanical effects of electron motion, characterized by the *Bohr radius* $r_B \approx 0.5 \times 10^{-10}$ m – see, e.g., QM Eq. (1.13).

⁵ Here e denotes the positive fundamental charge, $e \approx 1.6 \times 10^{-19}$ C, so that the electron charge equals $(-e)$.

⁶ See, e.g., SM Sec. 3.1.

⁷ This equation and/or its straightforward generalization to the case of charged particles (ions) of several kinds is frequently (especially in the theories of electrolytes and plasmas) called the *Debye-Hückel equation*.

where constant λ in this case is equal to the so-called *Debye screening length* λ_D , defined by relation

$$\lambda_D^2 \equiv \frac{\epsilon_r \epsilon_0 k_B T}{e^2 n}. \quad (2.8)$$

Debye
screening
length

Equation (7) is easy to solve: it describes an exponential decrease of the electric potential, with the characteristic length λ_D : $\phi \propto \exp\{-z/\lambda_D\}$. Plugging in the fundamental constants ϵ_0 , e , and k_B , we get the following estimate: $\lambda_D[\text{m}] \approx 70 (\epsilon_r T[\text{K}]/n[\text{m}^{-3}])^{1/2}$. According to this formula, in semiconductors at room temperature, the Debye length may be rather substantial. For example, in silicon ($\epsilon_r \approx 12$) doped to the charge carrier concentration $n = 3 \times 10^{24} \text{ m}^{-3}$ (the value typical for modern integrated circuits),⁸ $\lambda_D \approx 2 \text{ nm}$, still well above the atomic size scale a_0 . However, for typical good metals ($n \sim 10^{28} \text{ m}^{-3}$, $\epsilon_r \sim 10$) the same formula gives an estimate $\lambda_D \sim 4 \times 10^{-11} \text{ m}$, less than a_0 . In this case Eq. (8) should not be taken too literally, because it is based on the assumption of continuous charge distribution.

(iii) *Quantum statistics*. Actually, the last estimate is not valid for good metals (and highly doped semiconductors) for one more reason: their free electrons obey quantum (*Fermi-Dirac*) statistics rather than the Boltzmann distribution (4).⁹ As a result, at all realistic temperatures they form a degenerate quantum gas, occupying all available energy states below certain level $\mathcal{E}_F \gg k_B T$ called the *Fermi energy*. In these conditions, the screening of relatively low electric field¹⁰ may be described by replacing Eq. (5) with

$$\rho = e(n - n_e) = -eg(\mathcal{E}_F)(-U) = -e^2 g(\mathcal{E}_F)\phi, \quad (2.9)$$

where $g(\mathcal{E})$ is the density of quantum states (per unit volume) at electron's energy \mathcal{E} . At the Fermi surface, the density is of the order of n/\mathcal{E}_F .¹¹ As a result, we again get the second of Eqs. (7), but with a different characteristic scale λ , defined by the following relation:

$$\lambda_{\text{TF}}^2 \equiv \frac{\epsilon_r \epsilon_0}{e^2 g(\mathcal{E}_F)} \sim \frac{\epsilon_r \epsilon_0 \mathcal{E}_F}{e^2 n}, \quad (2.10)$$

Thomas-
Fermi
screening
length

and called the *Thomas-Fermi screening length*. Since for most good metals, n is of the order of 10^{29} m^{-3} , and \mathcal{E}_F is of the order of 10 eV, Eq. (10) typically gives λ_{TF} close to a few a_0 , and makes the Thomas-Fermi screening theory valid at least semi-quantitatively.

To summarize, the electric field penetration into good conductors is limited to a depth λ ranging from fractions of a nanometer to a few nanometers, so that for problems with the characteristic size much larger than that scale, the macroscopic boundary conditions (1) give a very good accuracy, and we will use them in the rest of this chapter. However, the reader should remember that in some situations

⁸ There is a good reason for making an estimate of λ_D for this case: the electric field created by the gate electrode of a field-effect transistor, penetrating into doped silicon by a depth $\sim \lambda_D$, controls current in this most important electronic device - on whose back all the current information revolution rides. Because of that, λ_D establishes the possible scale of semiconductor circuit shrinking which is the basis of the well-known Moore's law. (Practically, the scale is determined by integrated circuit patterning techniques, and Eq. (8) may be used to find the proper charge carrier density n and hence the level of silicon doping.)

⁹ See, e.g., SM Sec. 2.8. For a more detailed derivation of Eq. (10), see SM Chapter 3.

¹⁰ In good metals this equation is valid up to the fields $\sim E_F/e\lambda_{\text{TF}} \sim 10^9 \text{ V/m}$, very high by the usual standards. For example, the electric breakdown threshold for vacuum (or air-filled) gaps is $\sim 3 \times 10^6 \text{ V/m}$.

¹¹ See, e.g., SM Sec. 3.3.

involving semiconductors, as well as at nanoscale experiments with metals, the electric field penetration effect should be taken into account.

2.2. Capacitance

Let us start with systems consisting of charged conductors alone. Our goal here is calculating the distributions of electric field \mathbf{E} and potential ϕ in space, and the distribution of the surface charge density σ over the conductor surfaces. However, before doing that for particular situations, let us see if there are any integral measures of these distributions, that should be our primary focus.

The simplest case is of course a single conductor in the otherwise free space. According to Eq. (1), all its volume should have a constant electrostatic potential ϕ , evidently providing one convenient global measure of the situation. Another integral measure is evidently provided by the total charge

$$Q \equiv \int_V \rho d^3r = \oint_S \sigma d^2r, \quad (2.11)$$

where the latter integral is extended over the whole surface S of the conductor. In the general case, what we can tell about the relation between Q and ϕ ? At $Q = 0$, there is no electric field in the system, and it is natural (though not necessary) to select the arbitrary constant in the electrostatic potential to have $\phi = 0$. Then, if the conductor is charged with a finite Q , according to the Coulomb law, the electric field in any point of space is proportional to Q . Hence the electrostatic potential everywhere, including its value ϕ on the conductor, is also proportional to Q :

$$\phi = pQ. \quad (2.12)$$

The proportionality coefficient p , that depends on the conductor size and shape but not on Q , is called the *reciprocal capacitance* (or, not too often, “electrical elastance”). Usually, Eq. (12) is rewritten in a different form,

Self-
capacitance

$$Q = C\phi, \quad C \equiv \frac{1}{p}, \quad (2.13)$$

where C is called *self-capacitance*. (Frequently, C is called just *capacitance*, but we will soon see that for more complex situations the latter term may be too ambiguous.)

Before going to calculation of C , let us have a look at the electrostatic energy of a single conductor. In order to calculate it, of the several equations discussed in Chapter 1, Eq. (1.63) is most convenient, because all elementary charges q_k are now parts of the conductor surface charge, and hence sit at the same potential ϕ . As a result, the equation becomes very simple:

$$U = \frac{1}{2} \phi \sum_k q_k = \frac{1}{2} \phi Q. \quad (2.14)$$

Moreover, using the linear relation (13), the same result may be re-written in two more forms:

Electro-
static
energy

$$U = \frac{Q^2}{2C} = \frac{C}{2} \phi^2. \quad (2.15)$$

We will discuss several ways to calculate C in the next sections, and right now will have a quick look at just the simplest example for which we have calculated everything necessary in the previous

chapter: a conducting sphere of radius R . Indeed, we already know the electric field distribution: according to Eq. (1), $E = 0$ inside the sphere, while Eq. (1.19), with $Q(r) = Q$, describes the field distribution outside it. Moreover, since the latter formula is exactly the same as for the point charge placed in the sphere's center, the potential distribution in space can be obtained from Eq. (1.35) by replacing q with sphere's full charge Q . Hence, on the surface of the sphere (and, according to Eq. (2), through its interior),

$$\phi = \frac{1}{4\pi\epsilon_0} \frac{Q}{R}. \quad (2.16)$$

Comparing this result with the definition (13), for the self-capacitance we obtain¹²

$$C = 4\pi\epsilon_0 R = 2\pi\epsilon_0 D, \quad D \equiv 2R. \quad (2.17)$$

This formula, which should be well familiar to the reader, is convenient to get some feeling of how large the SI unit of capacitance (1 *farad*, abbreviated as F) is: the self-capacitance of Earth ($R_E \approx 6.34 \times 10^6$ m) is below 1 mF! Another important note is that while Eq. (17) is not exactly valid for a conductor of arbitrary shape, it implies an important estimate

$$C \sim 2\pi\epsilon_0 a \quad (2.18)$$

where a is the scale of the linear size of any conductor.¹³

Now proceeding to a system of two conductors, we immediately see why we should be careful with the capacitance definition: one constant C is insufficient to describe such system. Indeed, here we have two, generally different conductor potentials, ϕ_1 and ϕ_2 , that may depend on both conductor charges, Q_1 and Q_2 . Using the same arguments as for the one-conductor case, we may conclude that the dependence is always linear:

$$\begin{aligned} \phi_1 &= p_{11}Q_1 + p_{12}Q_2, \\ \phi_2 &= p_{21}Q_1 + p_{22}Q_2, \end{aligned} \quad (2.19)$$

but still has to be described not with one but with four coefficients $p_{jj'}$ ($j, j' = 1, 2$) forming the so-called *reciprocal capacitance matrix*

$$\begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix}. \quad (2.20)$$

Plugging relation (19) into Eq. (1.63), we see that the full electrostatic energy of the system may be expressed by a quadratic form:

¹² In the Gaussian units, using the standard replacement $4\pi\epsilon_0 \rightarrow 1$, this relation takes a remarkably simple form: $C = R$, good to remember. Generally, in the Gaussian units (but not in the SI system!) the capacitance has the dimensionality of length, i.e. is measured in centimeters. Note also that a convenient fractional SI unit, 1 picofarad (10^{-12} F) is very close to the Gaussian unit: $1 \text{ pF} = (1 \times 10^{-12}) / (4\pi\epsilon_0 \times 10^{-2}) \approx 0.8998 \text{ cm}$.

¹³ These arguments are somewhat insufficient to say which size should be used for a in the case of narrow, extended conductors, e.g., a thin, long wire of length L and diameter $D \ll L$. In the Very soon we will see that in such cases the electrostatic energy, and hence C , should mostly depend on the *larger* size of the conductor.

$$U = \frac{p_{11}}{2} Q_1^2 + \frac{p_{12} + p_{21}}{2} Q_1 Q_2 + \frac{p_{22}}{2} Q_2^2. \quad (2.21)$$

It is evident that the middle term in the right-hand part of this equation describes the electrostatic coupling of the conductors. (Without it, the energy would be just a sum of two independent electrostatic energies of conductors 1 and 2.) This is why systems with $|p_{12}|, |p_{21}| \ll p_{11}, p_{22}$ are called *weakly coupled*, and may be analyzed using approximate methods – see, e.g., Fig. 3 and its discussion below.

Before proceeding further, let us use the Lagrangian formalism of analytical mechanics¹⁴ to argue that the off-diagonal elements of matrix p_{jj} are always equal:

$$p_{12} = p_{21}. \quad (2.22)$$

Indeed, charges $Q_{1,2}$ may be taken for generalized coordinates q_j ($j = 1, 2$) of the system; then the corresponding generalized forces may be found as

$$\mathcal{F}_j = -\frac{\partial U}{\partial q_j} = -\frac{\partial U}{\partial Q_j}. \quad (2.23)$$

Applying this equation to Eq. (21), we see that, for example

$$\mathcal{F}_1 = -\left(p_{11}Q_1 + \frac{p_{12} + p_{21}}{2}Q_2\right). \quad (2.24)$$

Now we may argue that dynamics of charge Q_j should only depend on the electrostatic potential ϕ_j this charge “sees”. This means, in particular, that ϕ_1 should be a unique function of \mathcal{F}_1 . Comparing Eq. (24) with the first of Eqs. (19), we see that for this to be true, Eq. (22) should indeed be valid.

Equations (19) and (21) show that for the general case of arbitrary charges Q_1 and Q_2 , the system properties cannot be reduced to just one coefficient (“capacitance”). Let us consider three particular cases when such a reduction is possible.

(i) The system as the whole is electrically neutral: $Q_1 = -Q_2 \equiv Q$. In this case the most important function of Q is the difference of conductor potentials, called *voltage*:¹⁵

Voltage

$$V \equiv \phi_1 - \phi_2, \quad (2.25)$$

For that function, the subtraction of two Eqs. (19) gives

Mutual capacitance

$$V = \frac{Q}{C_m}, \quad \text{with } C_m \equiv \frac{1}{(p_{11} + p_{22}) - (p_{12} + p_{21})}, \quad (2.26)$$

where coefficient C_m is called the *mutual capacitance* between the conductors – or, again, just “capacitance”. The same coefficient describes the electrostatic energy of the system. Indeed, plugging Eq. (25) into Eq. (21), we see that both forms of Eq. (15) are reproduced if ϕ is replaced with V , Q_1 with Q , and C with C_m :

¹⁴ See, e.g., CM Chapter 2.

¹⁵ A word of caution: in condensed matter physics, voltage is usually defined differently, as the difference of *electrochemical* rather than *electrostatic* potentials – see, e.g., SM Sec. 6.4. These two definitions coincide if the conductors have equal *workfunctions* (for example, if they are made of the same material), and in this course their difference will be ignored.

$$U = \frac{Q^2}{2C_m} = \frac{C_m}{2} V^2. \quad (2.27)$$

Capacitor's energy

The best known system for which the mutual capacitance C_m may be readily calculated is the *plane* (or “parallel-plate”) *capacitor*, a system of two conductors separated with a narrow, plane gap (Fig. 2). Indeed, since the surface charges, that contribute to the opposite charges $\pm Q$ of the conductors in this system, attract each other, in the limit $d \ll a$ they sit entirely on the sides of the narrow gap.

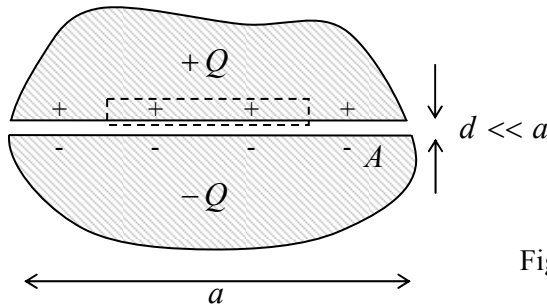


Fig. 2.2. Plane capacitor.

Let us apply the Gauss law to a pillbox volume (shown by dashed line in Fig. 2) whose area is a small part of the gap (but nevertheless much larger than d^2), with one of the plane lids inside a conductor, and another one inside the gap. The result immediately shows that the electric field within the gap is $E = \sigma/\epsilon_0$, i.e. is independent of the pillbox thickness. Integrating this field across thickness d of the gap, we get $V = Ed = \sigma d/\epsilon_0$, so that $\sigma = \epsilon_0 V/d$. But this voltage should not depend on the selection of the point of the gap area. As a result, σ should be also constant over all the gap area A , and hence $Q = \sigma A = \epsilon_0 V/d$. Thus we may write $V = Q/C_m$, with

$$C_m = \frac{\epsilon_0}{d} A. \quad (2.28)$$

 C_m of planar capacitor

Let me offer a few comments on this well-known formula. First, it is valid even if the gap is not quite planar, for example if it gently curves on a scale much larger than d . Second, Eq. (28) is only valid if $A \sim a^2$ is much larger than d^2 , because its derivation ignores the electric field deviations from uniformity¹⁶ at distances $\sim d$ near the gap edges. Finally, the same condition ($A \gg d^2$) assures that C_m is much larger than the self-capacitance of each of the conductors – see Eq. (18). The opportunities given by this fact for electronic engineering and experimental physics practice are rather astonishing. For example, a very realistic 3-nm layer of high-quality aluminum oxide (which may provide a nearly perfect electric insulation between two thin conducting films) with area of 0.1 m² (which is a typical area of silicon wafers used in semiconductor industry) provides $C_m \sim 1$ mF,¹⁷ larger than the self-capacitance of the whole planet Earth!

In the case shown in Fig. 2, the electrostatic coupling of the two conductors is evidently strong. As an opposite example of a weakly coupled system, let us consider two conducting spheres of the same radius R , separated by a much larger distance d (Fig. 3).

¹⁶ Frequently referred to “fringe” fields resulting in an additional “stray” capacitance $C_m' \sim \epsilon_0 a$.

¹⁷ Just as in Sec. 1, in order for the estimate to be realistic, I took into account the additional factor ϵ_r (for aluminum oxide, close to 10) which should be included into the nominator of Eq. (28) to make it applicable to dielectrics – see Chapter 3 below.

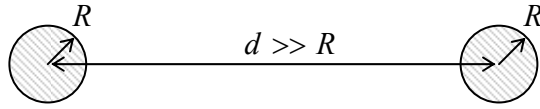


Fig. 2.3. A system of two well separated, similar conducting spheres.

In this case the diagonal components of matrix p_{ij} may be approximately found from Eq. (16), i.e. by neglecting the coupling altogether:

$$p_{11} = p_{22} \approx \frac{1}{4\pi\epsilon_0 R}. \quad (2.29)$$

Now, if we had just one sphere (say, number 1), the electric potential at distance d from its center would be given by Eq. (16): $\phi = Q_1/4\pi\epsilon_0 d$. Now if we move into this point a small ($R \ll d$) sphere without its own charge, we may expect that its potential should not be too far from this result, so that $\phi_2 \approx Q_1/4\pi\epsilon_0 d$. Comparing this expression with Eq. (19) (taken for $Q_2 = 0$), we get

$$p_{12} = p_{21} \approx \frac{1}{4\pi\epsilon_0 d} \ll p_{11}, p_{22}. \quad (2.30)$$

From here and Eq. (26), the mutual capacitance

$$C_m \approx \frac{1}{p_{11} + p_{22}} \approx 2\pi\epsilon_0 R. \quad (2.31)$$

We see that (somewhat counter-intuitively), in this case C_m does not depend substantially on the distance between the spheres, i.e. does *not* describe their electrostatic coupling. The off-diagonal coefficients of the reciprocal capacitance matrix (20) play this role much better – see Eq. (30).

(ii) Now let us consider the case when only one conductor of the two is charged, for example $Q_1 \equiv Q$, while $Q_2 = 0$. Then Eqs. (19) yield

$$\phi_1 = p_{11}Q_1. \quad (2.32)$$

Now, if we follow Eq. (13) and define $C_j \equiv 1/p_{jj}$ as the *partial capacitance* of conductor number j , we see that it differs from the mutual capacitance C_m – cf. Eq. (26). For example, in the case shown in Fig. 3, $C_1 = C_2 \approx 4\pi\epsilon_0 R \approx 2C_m$.

(iii) Finally, let us consider a popular case when one of the conductors is charged by a certain charge (say, $Q_1 = Q$), but the potential of another one is sustained constant, say $\phi_2 = 0$.¹⁸ (This condition is especially easy to implement if the second conductor is much larger than the first one. Indeed, as the estimate (18) shows, in this case it would take much larger charge Q_2 to make potential ϕ_2 comparable with ϕ_1 .) In this case the second of equations (19) yields $Q_2 = -(p_{21}/p_{22})Q_1$. Plugging this relation into the first of those equations, we get

¹⁸ In electrical engineering, such constant-potential conductor is called the *ground*. This term stems from the fact that in many cases the Earth surface may be considered a good electric ground, because its potential is unaffected by laboratory-scale electric charges.

$$\phi_1 = \left(p_{11} - \frac{p_{12}p_{21}}{p_{22}} \right) Q_1 \quad (2.33)$$

Thus, if we treat the reciprocal of the expression in parentheses,

$$C_1^{\text{ef}} \equiv \left(p_{11} - \frac{p_{12}p_{21}}{p_{22}} \right)^{-1} \quad (2.34)$$

as the *effective capacitance* of the first conductor, it is generally different both from C_m and (unless the conductors are far apart and their electrostatic coupling is negligible) from $C_1 = 1/p_{11}$.

To summarize this section, the potential (and hence the actual capacitance) of a conductor in a two-conductor system may be very much dependent on what exactly is being done with the second conductor when the first one is charged. This is also true for multi-conductor systems (for whose description, Eqs. (19) and (21) may be readily generalized); moreover, in that case even the mutual capacitance between two selected conductors may depend on the electrostatics conditions of other components of the system.

2.3. The simplest boundary problems

In the general case when the electric field distribution in the free space between the conductors cannot be readily found from the Gauss law or by any other special methods, the best approach is to try to solve the differential Laplace equation (1.42), with boundary conditions (1b):

$$\nabla^2 \phi = 0, \quad \phi|_{S_k} = \phi_k, \quad (2.35)$$

Typical
boundary
problem

where S_k is the surface of the k -th conductor of the system. After such *boundary problem* has been solved, i.e. the spatial distribution $\phi(\mathbf{r})$ has been found in all points outside the conductor, it is straightforward to use Eq. (3) to find the surface charge density, and finally the total charge

$$Q_k = \oint_{S_k} \sigma d^2r \quad (2.36)$$

of each conductor, and hence any component of the reciprocal capacitance matrix p_{jj} . As an illustration, let us implement this program for three very simple problems.

(i) Plane capacitor (Fig. 2). In this case, the easiest way to solve the Laplace equation is to use linear (Cartesian) coordinates with one coordinate axis, say z , normal to the conductor surfaces (Fig. 4).

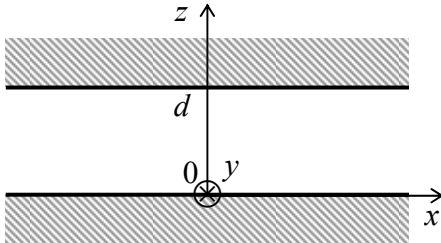


Fig. 2.4. Plane capacitor's geometry used for the solution of the boundary problem (35).

In these coordinates, the Laplace operator is just the sum of three second derivatives.¹⁹ It is evident that due to problem's translational symmetry in the $\{x, y\}$ plane, deep inside the gap (i.e. at the lateral distance from the edges much larger than d) the electrostatic potential may only depend on the coordinate perpendicular to the gap surfaces: $\phi(\mathbf{r}) = \phi(z)$. For such a function, derivatives over x and y vanish, and the boundary problem (35) is reduced to a very simple ordinary differential equation

$$\frac{d^2\phi}{dz^2}(z) = 0, \quad (2.37)$$

with boundary conditions

$$\phi(0) = 0, \quad \phi(d) = V. \quad (2.38)$$

(For the sake of notation simplicity, I have used the discretion of adding a constant to the potential to make one of the potentials vanish, and also definition (25) of voltage V .) The general solution of Eq. (37) is a linear function: $\phi(z) = c_1 z + c_2$, whose constant coefficients $c_{1,2}$ may be found, in an elementary way, from the boundary conditions (38). The final solution is

$$\phi = V \frac{z}{d}. \quad (2.39)$$

From here the only nonvanishing component of the electric field is

$$E_z = -\frac{d\phi}{dz} = -\frac{V}{d}, \quad (2.40)$$

and the surface charge of the capacitor plates

$$\sigma = \varepsilon_0 E_n = \mp \varepsilon_0 E_z = \pm \varepsilon_0 \frac{V}{d}, \quad (2.41)$$

where the upper and lower sign correspond to the upper and lower plate, respectively. Since σ does not depend on coordinates x and y , we can get the full charges $Q_1 = -Q_2 \equiv Q$ of the surfaces by its multiplication by the gap area A , giving us the again already known result (26) for the mutual capacitance $C_m \equiv Q/V$. I believe that this calculation, though very easy, may serve as a good introduction to the boundary problem solution philosophy.

(ii) Coaxial-cable capacitor. *Coaxial cable* is a system of two round cylindrical, coaxial conductors, with the cross-section shown in Fig. 5.

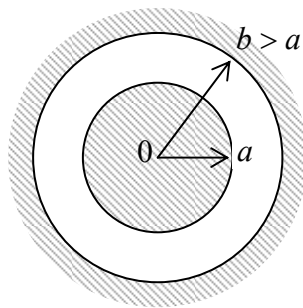


Fig. 2.5. Cross-section of a coaxial capacitor.

¹⁹ See, e.g. MA Eq. (9.1).

Evidently, in this case the cylindrical coordinates $\{\rho, \phi, z\}$, with axis z along the common axis of the cylinders, are most appropriate. Due to the axial symmetry of the problem, in these coordinates $\mathbf{E}(\mathbf{r}) = \mathbf{n}_\rho E(\rho)$, $\phi(\mathbf{r}) = \phi(\rho)$, so that in the general expression for the Laplace operator²⁰ we can take $\partial/\partial\phi = \partial/\partial z = 0$. As a result, only the first (radial) term of the operator survives, and the boundary problem (35) takes the form

$$\frac{1}{\rho} \frac{d}{d\rho} \left(\rho \frac{d\phi}{d\rho} \right) = 0, \quad \phi(a) = V, \quad \phi(b) = 0. \quad (2.42)$$

The sequential integration of this ordinary differential equation is elementary (and similar to that of the Poisson equation in spherical coordinates, performed in Sec. 1.3), giving

$$\frac{d\phi}{d\rho} = \frac{c_1}{\rho}, \quad \phi(\rho) = c_1 \int_a^\rho \frac{d\rho'}{\rho'} + c_2 = c_1 \ln \frac{\rho}{a} + c_2. \quad (2.43)$$

Constants $c_{1,2}$ may be found using boundary conditions (42):

$$V = c_2, \quad 0 = c_1 \ln \frac{b}{a} + c_2, \quad (2.44)$$

giving $c_1 = -V/\ln(b/a)$, so that solution (43) takes the following form

$$\phi(\rho) = V \left(1 - \frac{\ln(\rho/a)}{\ln(b/a)} \right). \quad (2.45)$$

Next, for our axial symmetry the general expression for the gradient²¹ is reduced to the radial derivative, so that

$$E(\rho) \equiv -\frac{d\phi(\rho)}{d\rho} = \frac{V}{\rho \ln(b/a)}. \quad (2.46)$$

This expression, plugged into Eq. (2), allows us to find the density of conductors' surface charge. For example, for the inner electrode

$$\sigma_a = \varepsilon_0 E_a = \frac{\varepsilon_0 V}{a \ln(b/a)}, \quad (2.47)$$

so that its full charge (per unit length of the system) is

$$\frac{Q}{L} = 2\pi a \sigma_a = \frac{2\pi \varepsilon_0 V}{\ln(b/a)}. \quad (2.48)$$

(It is straightforward to check that the charge of the outer electrode is equal and opposite.) Hence, by the definition of the mutual capacitance, its value per unit length is

$$\frac{C_m}{L} \equiv \frac{Q}{LV} = \frac{2\pi \varepsilon_0}{\ln(b/a)}. \quad (2.49)$$

²⁰ See, e.g., MA Eq. (10.3).

²¹ See, e.g., MA Eq. (10.2).

This expression shows that the total capacitance C is proportional to the systems length L (if $L \gg a, b$), while being only logarithmically dependent on is the dimensions of its cross-section. Since log of a very large argument is an extremely slow function (sometimes called “quasi-constant”), if the external conductor is made large ($b \gg a$) the capacitance diverges, but very weakly. Such a logarithmic divergence may be cut by any miniscule additional effect, for example by the finite length L of the system. This allows one to get a crude but very useful estimate of self-capacitance of a *single* wire:

$$C \approx \frac{2\pi\epsilon_0 L}{\ln(L/a)}, \quad \text{for } L \gg a. \quad (2.50)$$

On the other hand, if the gap between the conductors is narrow: $b = a + d$, with $d \ll a$, then $\ln(b/a) = \ln(1 + d/a)$ may be approximated as d/a , and Eq. (49) is reduced to $C_m \approx 2\pi\epsilon_0 aL/d$, i.e. to Eq. (28) for the plane capacitor, with $A = 2\pi aL$.

(iii) Spherical capacitor. This is a system of two conductors, with the same central cross-section as the coaxial cable (Fig. 5), but now with the spherical rather than axial symmetry. This symmetry implies that we are better off using spherical coordinates, so that potential ϕ depends only on one of them, the distance r from the common center of the conductors: $\phi(\mathbf{r}) = \phi(r)$. As we already know from Sec. 1.3, in this case the general expression for the Laplace operator is reduced to its first (radial) term, so that the Laplace equation takes a simple form – see Eq. (1.47). Moreover, we have already found the general solution to this equation – see Eq. (1.50):

$$\phi(r) = \frac{c_1}{r} + c_2, \quad (2.51)$$

Now acting exactly as above, i.e. determining constant c_1 from the boundary conditions $\phi(a) = V$, $\phi(b) = 0$, we get

$$V = c_1 \left(\frac{1}{a} - \frac{1}{b} \right), \quad \text{so that} \quad \phi(r) = \frac{V}{r} \left(\frac{1}{a} - \frac{1}{b} \right)^{-1} + c_2. \quad (2.52)$$

Next, we can use the spherical symmetry to find electric field, $\mathbf{E}(\mathbf{r}) = \mathbf{n}_r E(r)$, with

$$E(r) = -\frac{d\phi}{dr} = \frac{V}{r^2} \left(\frac{1}{a} - \frac{1}{b} \right)^{-1}, \quad (2.53)$$

and hence its values on conductors' surfaces, and then the surface charge density σ from Eq. (2). For example, for the inner conductor's surface,

$$\sigma_a = \epsilon_0 E(a) = \epsilon_0 \frac{V}{a^2} \left(\frac{1}{a} - \frac{1}{b} \right)^{-1}, \quad (2.54)$$

so that, finally, for the full charge of that conductor we get

$$Q = 4\pi a^2 \sigma = 4\pi\epsilon_0 \left(\frac{1}{a} - \frac{1}{b} \right)^{-1} V. \quad (2.55)$$

(Again, the charge of the outer conductor is equal and opposite.) Now we can use the definition of the mutual capacitance to get the final result

$$C_m \equiv \frac{Q}{V} = 4\pi\epsilon_0 \left(\frac{1}{a} - \frac{1}{b} \right)^{-1} = 4\pi\epsilon_0 \frac{ab}{b-a}. \quad (2.56)$$

For $b \gg a$, this result coincides with Eq. (17) for self-capacitance of the inner conductor. On the other hand, if the gap between two conductors is narrow, $d \equiv b - a \ll a$,

$$C_m = 4\pi\epsilon_0 \frac{a(a+d)}{d} \approx 4\pi\epsilon_0 \frac{a^2}{d}, \quad (2.57)$$

i.e. the capacitance approaches that of the planar capacitor of area $A = 4\pi a^2$ - as it should.

All this seems (and is) very straightforward, but let us contemplate what was the reason for such easy successes. We have managed to find such coordinate transformations, for example $\{x, y, z\} \rightarrow \{r, \theta, \varphi\}$ in the spherical case, that both the Laplace equation and the boundary conditions involve only one of the new coordinates (in this case, r). The necessary condition for the former fact is that the new coordinates (in this case, spherical ones) are *orthogonal*. This means that three vector components of differential $d\mathbf{r}$, due to small variations of the new coordinates (say, dr , $d\theta$, and $d\varphi$), are mutually perpendicular. If this were not so, the Laplace operator would not fall into the simple sum of three independent parts, and could not be reduced, at the proper symmetry of the problem, to just one of these components, making it readily integrable.

2.4. Other orthogonal coordinates

Since the cylindrical and spherical coordinates are only simplest examples of the orthogonal (or “orthogonal curvilinear”) coordinates, this methodology may be extended to other coordinate systems of this type. As an example, let us have a look at the following problem: finding the self-capacitance of a thin, round conducting disk (and, as solution’s by-products, the distributions of the electric field and surface charge) – see Fig. 6. The cylindrical or spherical coordinates would not give too much help here, because though they have the appropriate axial symmetry about axis z , they would make the boundary condition on the disk too complex (two coordinates, either ρ and z , or r and θ).

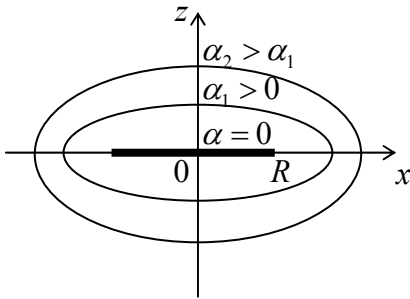


Fig. 2.6. The thin conducting disk problem. (The cross-section of the system by the vertical plane $y = 0$.)

The relief comes from noting that the disk, i.e. the area $z = 0$, $r < R$, may be thought of as the limiting case of an *axially-symmetric ellipsoid* - the result of rotation of the usual ellipse about one of its axes - in our case, the vertical axis z .²² Analytically, such an ellipsoid may be described by the following equation:

²² Alternative names for this surface are “degenerate ellipsoid”, “ellipsoid of rotation”, and “spheroid”.

$$\frac{x^2 + y^2}{a^2} + \frac{z^2}{b^2} = 1, \quad (2.58)$$

where a and b are the so-called *major semi-axes* whose ratio determines the ellipse *eccentricity* (the degree of squeezing). For our problem, we will only need *oblate* ellipsoids with $a \geq b$; according to Eq. (58), they may be presented as surfaces of constant α in the system of *degenerate ellipsoidal* (or “spheroidal”) *coordinates* $\{\alpha, \beta, \varphi\}$, which are related to the Cartesian coordinates as follows:

$$\begin{aligned} x &= R \cosh \alpha \sin \beta \cos \varphi, \\ y &= R \cosh \alpha \sin \beta \sin \varphi, \\ z &= R \sinh \alpha \cos \beta. \end{aligned} \quad (2.59)$$

Such ellipsoidal coordinates are the evident generalization of the spherical coordinates, which correspond to the limit $\alpha \gg 1$ (i.e. $r \gg R$). In the opposite limit of small α , the surface of constant $\alpha = 0$ describes our thin disk of radius R . It is almost evident (and easy to prove) that coordinates (59) are also orthogonal, so that the Laplace operator may be expressed as a sum of three independent terms:

$$\nabla^2 = \frac{1}{R^2 (\cosh^2 \alpha - \sin^2 \beta)} \times \left[\frac{1}{\cosh \alpha} \frac{\partial}{\partial \alpha} \left(\cosh \alpha \frac{\partial}{\partial \alpha} \right) + \frac{1}{\sin \beta} \frac{\partial}{\partial \beta} \left(\sin \beta \frac{\partial}{\partial \beta} \right) + \left(\frac{1}{\sin^2 \beta} - \frac{1}{\cosh^2 \alpha} \right) \frac{\partial^2}{\partial \varphi^2} \right]. \quad (2.60)$$

Though this expression may look a bit intimidating, let us notice that in our current problem, the boundary conditions depend only on coordinate α :²³

$$\phi|_{\alpha=0} = V, \quad \phi|_{\alpha=\infty} = 0. \quad (2.61)$$

Hence there is every reason to believe that the electrostatic potential in all space is the function of α alone. (In other words, all ellipsoids $\alpha = \text{const}$ are the equipotential surfaces.) Indeed, acting on such function $\phi(\alpha)$ by the Laplace operator (60), we see that the two last terms in the square brackets vanish, and the Laplace equation (35) is reduced to a simple ordinary differential equation

$$\frac{d}{d\alpha} \left[\cosh \alpha \frac{d\phi}{d\alpha} \right] = 0. \quad (2.62)$$

Integrating it twice, just as we did in the previous problems, we get

$$\phi(\alpha) = c_1 \int \frac{d\alpha}{\cosh \alpha}. \quad (2.63)$$

This integral may be readily taken, for example, using the substitution $\xi \equiv \sinh \alpha$ (with $d\xi \equiv \cosh \alpha d\alpha$, $\cosh^2 \alpha = 1 + \sinh^2 \alpha = 1 + \xi^2$):

$$\phi(\alpha) = c_1 \int_0^{\sinh \alpha} \frac{d\xi}{1 + \xi^2} + c_2 = c_1 \arctan(\sinh \alpha) + c_2. \quad (2.64)$$

²³ I have called disk's potential V , to distinguish it from the potential ϕ at an arbitrary point of space.

The integration constants $c_{1,2}$ are again simply found from boundary conditions, in this case Eqs. (61), and we arrive at the final expression for the electrostatic potential:

$$\phi(\alpha) = V \left[1 - \frac{2}{\pi} \arctan(\sinh \alpha) \right]. \quad (2.65)$$

This solution satisfies both the Laplace equation and the boundary conditions. Mathematicians tell us that the solution of any boundary problem of the type (35) is *unique*, so we do not need to look any further.

Now we may use Eq. (2) to find the surface density of electric charge, but in the case of thin disk, it is more natural to add up such densities on its top and bottom surfaces at the same distance $r = (x^2 + y^2)^{1/2}$ from the disk center (which are evidently equal, due to the problem symmetry about plane $z = 0$): $\sigma = 2\epsilon_0 E_n|_{z=+0}$. According to Eq. (65), the electric field on the surface is

$$E_n|_{\alpha=+0} = -\frac{\partial \phi}{\partial z}\bigg|_{z=+0} = -\frac{\partial \phi(\alpha)}{\partial (R \sinh \alpha \cos \beta)}\bigg|_{\alpha=+0} = \frac{2}{\pi} V \frac{1}{R \cos \beta} = \frac{2}{\pi} V \frac{1}{(R^2 - r^2)^{1/2}}, \quad (2.66)$$

and we see that the charge is distributed along the disk very nonuniformly:

$$\sigma = \frac{4}{\pi} \epsilon_0 V \frac{1}{(R^2 - r^2)^{1/2}}, \quad (2.67)$$

with a singularity at the disk edge. Below we will see that such singularities are very typical for sharp edges of conductors.²⁴ Fortunately, in our current case the divergence is integrable, giving a finite disk charge:

$$Q = \int_{\text{disk surface}} \sigma d^2 r = \int_0^R \sigma(r) 2\pi r dr = \frac{4}{\pi} \epsilon_0 V 2\pi \int_0^R \frac{r dr}{(R^2 - r^2)^{1/2}} = 4\epsilon_0 V R \int_0^1 \frac{d\xi}{\sqrt{1-\xi}} = 8\epsilon_0 R V. \quad (2.68)$$

Thus, for disk's self-capacitance we get a very simple result,

$$C = 8\epsilon_0 R = \frac{2}{\pi} 4\pi\epsilon_0 R, \quad (2.69)$$

a factor of $2/\pi \approx 0.64$ lower than that for the conducting sphere of the same equal radius, but still complying with the general estimate (18).

Can we always find a “good” system of orthogonal coordinates? Unfortunately, the answer is *no*, even for highly symmetric geometries. This is why the practical value of this approach is limited, and other methods of boundary problems are clearly needed. Before moving to them, however, let us note that in the case of 2D problems (i.e. cylindrical geometries), the orthogonal coordinate method gets help from the following *conformal mapping* approach.

Let us consider the pair of Cartesian coordinates $\{x, y\}$ of the cross-section plane as a complex variable $z = x + iy$,²⁵ where i is the imaginary unity ($i^2 = -1$), and let $u(z) = u + iv$ be an *analytic complex*

²⁴ If you seriously worry about the formal infinity of charge density at $r \rightarrow R$, please remember that this mathematical artifact disappears for any nonvanishing disk thickness.

²⁵ The complex variable z should not be confused with the (real) 3rd spatial coordinate z ! We are considering 2D problems now, with the potential independent of z .

function of z .²⁶ For our current purposes, the most important property of an analytic function is that its real and imaginary parts obey the following *Cauchy-Riemann relations*:²⁷

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \quad \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}. \quad (2.70)$$

For example, for the function

$$w = z^2 = (x + iy)^2 = (x^2 - y^2) + 2ixy, \quad (2.71)$$

whose real and imaginary parts are

$$u \equiv \text{Re } w = x^2 - y^2, \quad v \equiv \text{Im } w = 2xy, \quad (2.72)$$

we immediately see that $\partial u/\partial x = 2x = \partial v/\partial y$, and $\partial v/\partial x = 2y = -\partial u/\partial y$, in accordance with Eq. (70).

Let us differentiate the first of Eqs. (70) over x again, then change the order of differentiation, and after that use the latter of those equations:

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial}{\partial x} \frac{\partial u}{\partial x} = \frac{\partial}{\partial x} \frac{\partial v}{\partial y} = \frac{\partial}{\partial y} \frac{\partial v}{\partial x} = -\frac{\partial}{\partial y} \frac{\partial u}{\partial y} = -\frac{\partial^2 u}{\partial y^2}, \quad (2.73)$$

and similarly for v . This means that the sum of second-order partial derivatives of each of real functions $u(x,y)$ and $v(x,y)$ is zero, i.e. that both functions obey the 2D Laplace equation. This mathematical fact opens a nice way of solving problems of electrostatics for (relatively simple) 2D geometries. Imagine that for a particular boundary problem we have found a function $w(z)$ for which either $u(x, y)$ or $v(x, y)$ is constant on all electrode surfaces. Then all lines of constant u (or v) present equipotential surfaces, i.e. the problem of the potential distribution has been essentially solved.

As a simple example, consider a practically important problem: the *quadrupole electrostatic lens*- a system of four cylindrical²⁸ electrodes with hyperbolic cross-sections, whose boundaries obey the following relations:

$$x^2 - y^2 = \begin{cases} +a^2, & \text{for the left and right electrodes,} \\ -a^2, & \text{for the top and bottom electrodes,} \end{cases} \quad (2.74)$$

voltage-biased as shown in Fig. 7a. Comparing these relations with Eqs. (72), we see that each electrode surface corresponds to a constant value of $u = \pm a^2$. Moreover, potentials of both surfaces with $u = +a^2$ are equal to $+V/2$, while those with $u = -a^2$ are equal to $-V/2$. Hence we may conjecture that the electrostatic potential at each point is a function of u alone; moreover, a simple linear function,

$$\phi = c_1 u + c_2 = c_1 (x^2 - y^2) + c_2, \quad (2.75)$$

²⁶ The analytic (or “holomorphic”) function may be defined as the one that may be expanded into the complex Taylor series, i.e. is infinitely differentiable in the given point. (Almost all “regular” functions, such as z^n , $z^{1/n}$, $\exp z$, $\ln z$, etc. and their combinations are analytic at all z , maybe besides certain special points.) If the reader needs to brush up his or her background on this subject, I can recommend a popular (and very inexpensive :-)) textbook by M. Spiegel *et al.*, *Complex Variables*, 2nd ed., McGraw-Hill, 2009.

²⁷ These relations may be, in particular, to prove the famous Cauchy integral formula – see, e.g., MA Eq. (15.1).

²⁸ Let me remind the reader that in mathematics, term *cylindrical* describes a surface formed by translation, along a straight line, of an arbitrary curve, and hence more general than the usual circular cylinder. (In this terminology, for example, a prism is also a particular form of cylinder, formed by translating a polygon.)

is a valid (and hence the unique) solution of our boundary problem. Indeed, it does satisfy the Laplace equation, while its constants $c_{1,2}$ may be selected in a way to satisfy all the boundary conditions shown in Fig. 7a:

$$\phi = \frac{V}{2} \frac{x^2 - y^2}{a^2}. \quad (2.76)$$

so that the boundary problem has been solved.

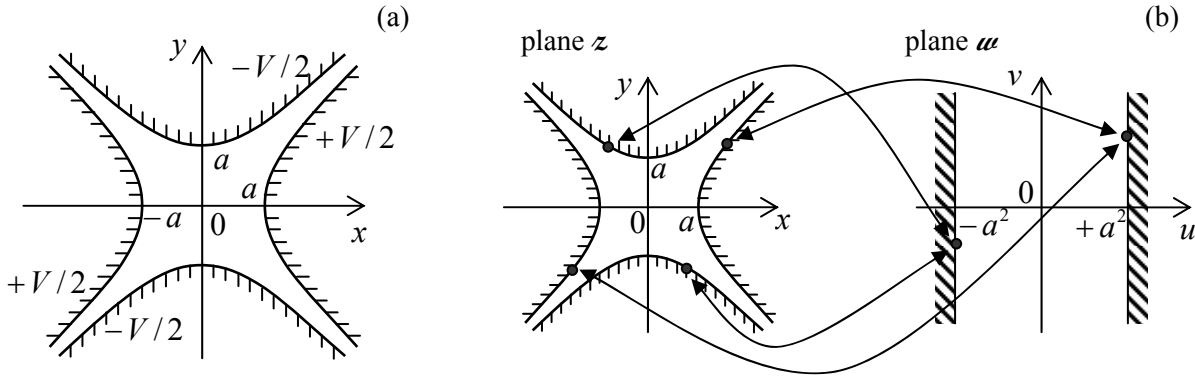


Fig. 2.7. (a) Quadrupole electrostatic lens geometry and (b) its analysis using conformal mapping.

According to Eq. (76), all equipotential surfaces are hyperbolic cylinders, similar to those of the electrode surfaces. What remains is to find the electric field at an arbitrary point inside the system:

$$E_x = -\frac{\partial \phi}{\partial x} = -V \frac{x}{a^2}, \quad E_y = -\frac{\partial \phi}{\partial y} = V \frac{y}{a^2}. \quad (2.77)$$

These formulas show that if charged particles (e.g., electrons in an electron optics system) are launched to fly ballistically through the lens, along axis z , they experience a force pushing them toward the symmetry axis and proportional to particle's deviation from the axis (and thus equivalent in action to an optical lens with positive refraction power) in one direction, and a force pushing them out (negative refractive power) in the perpendicular direction. One can show that letting charged particles fly through several such lenses, with alternating voltage polarities, in series, enables beam focusing.²⁹

Hence, we have reduced the 2D Laplace boundary problem to that of finding the proper analytic function $w(z)$. This task may be also understood as that of finding a *conformal map*, i.e. a correspondence between components of any point pair, $\{x, y\}$ and $\{u, v\}$, residing, respectively, on the initial Cartesian plane z and the plane w of the new variables. For example, Eq. (74) maps the real electrode configuration onto the plane capacitor with infinite area (Fig. 7b), and the simplicity of Eq. (75) is due to the fact that for the latter system the equipotential surfaces are just parallel planes.

For more complex geometries, the suitable analytic function $w(z)$ may be hard to find. However, for conductors with piece-linear cross-section boundaries, substantial help may be obtained from the following *Schwarz-Christoffel integral*

²⁹ See, e.g., textbook by P. Grivet, *Electron Optics*, 2nd ed., Pergamon, 1972, or the review collection A. Septier (ed.), *Focusing Charged Particles*, vol. I, Academic Press, 1967, in particular the review by K.-J. Hanszen and R. Lauer, pp. 251-307.

$$\omega(z) = \text{const} \times \int \frac{dz}{(z-x_1)^{k_1}(z-x_2)^{k_2} \dots (z-x_{N-1})^{k_{N-1}}} \quad (2.78)$$

that provides the conformal mapping of the interior of an arbitrary N -sided polygon on plane $\omega = u + iv$, and the upper-half ($y > 0$) of plane $z = x + iy$. Here x_j ($j = 1, 2, N-1$) are the points of axis $y = 0$ (i.e., of the boundary of the mapped region on plane z) to which the corresponding polygon vertices are mapped, while k_j are the exterior angles at the polygon vertices, measured in the units of π , with $-1 \leq k_j \leq +1$ – see Fig. 8.³⁰ Of points x_j , two may be selected arbitrarily (because their effects may be compensated by the multiplicative constant in Eq. (78), and the constant of integration), while all the others have to be adjusted to provide the correct mapping.

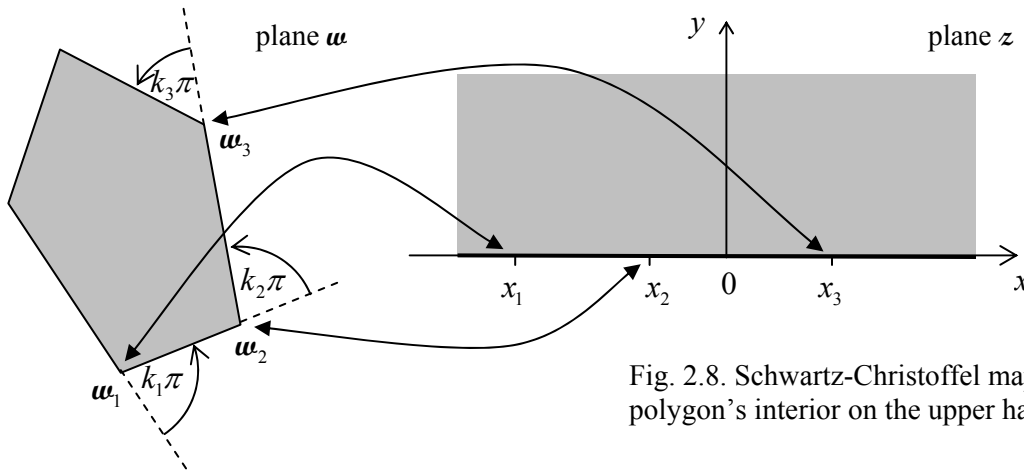


Fig. 2.8. Schwartz-Christoffel mapping of polygon's interior on the upper half-plane.

In the general case, the complex integral (78) may be hard to tackle. However, in some important cases, in particular those with right angles ($k_j = \pm 1/2$) and/or with some points ω_j at infinity, the integrals may be readily worked out, giving explicit analytical expressions for the mapping functions $\omega(z)$. For example, let us consider a semi-infinite strip, defined by restrictions $-1 \leq u \leq +1$ and $0 \leq v$, on plane ω – see the left panel of Fig. 9.

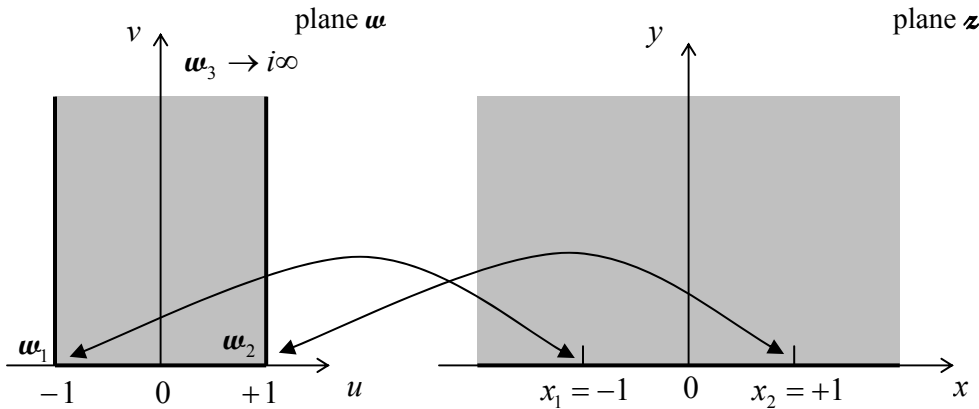


Fig. 2.9. Semi-infinite strip mapped onto the upper half-plane.

³⁰ Integral (70) includes only $(N-1)$ rather than N poles, because polygon's shape is completely determined by $(N-1)$ positions ω_j of its vertices and $(N-1)$ angles πk_j . In particular, since the algebraic sum of all external angles of a polygon equals π , the last angle parameter $k_j = k_N$ is uniquely determined by the set of the previous ones.

The strip may be considered as a polygon, with one vertex at the infinitely distant vertical point $w_3 = 0 + i\infty$. Let us map it on the upper half of plane z , shown on the right panel of Fig. 9, with vertex $w_1 = -1 + i0$ mapped onto point $x_1 = -1, y_1 = 0$, and vertex $w_2 = +1 + i0$ mapped onto point $x_2 = +1, y_2 = 0$. Since in this case both external angles are equal to $+\pi/2$, and hence $k_1 = k_2 = +1/2$, Eq. (78) is reduced to

$$w(z) = \text{const} \times \int \frac{dz}{(z+1)^{1/2}(z-1)^{1/2}} = \text{const} \times \int \frac{dz}{(z^2-1)^{1/2}} = \text{const} \times i \int \frac{dz}{(1-z^2)^{1/2}}. \quad (2.79)$$

This complex integral may be taken, just as for real z , by the substitution $z = \sin \xi$, giving

$$w(z) = \text{const}' \times \int_{\arcsin z}^{\arcsin z} d\xi = c_1 \arcsin z + c_2. \quad (2.80)$$

Determining constants $c_{1,2}$ from the required mapping, i.e. from the equations $w(-1 + i0) = -1 + i0$ and $w(+1 + i0) = +1 + i0$ (see Fig. 9), we finally get

$$w(z) = \frac{2}{\pi} \arcsin z, \quad \text{i.e. } z = \sin \frac{\pi w}{2}. \quad (2.81a)$$

Using the well-known expression for the sine of a complex argument,³¹ we may rewrite this elegant result in either of the two following forms for the real and imaginary components of z and w :

$$u = \frac{2}{\pi} \arcsin \frac{2x}{[(x+1)^2 + y^2]^{1/2} + [(x-1)^2 + y^2]^{1/2}}, \quad v = \frac{2}{\pi} \operatorname{arccosh} \frac{[(x+1)^2 + y^2]^{1/2} + [(x-1)^2 + y^2]^{1/2}}{2},$$

$$x = \sin \frac{\pi u}{2} \cosh \frac{\pi v}{2}, \quad y = \cos \frac{\pi u}{2} \sinh \frac{\pi v}{2}. \quad (2.81b)$$

It is amazing how perfectly does the last formula manage to keep $y \equiv 0$ at different borders of our w -region (Fig. 9): at its side borders ($u = \pm 1, 0 \leq v < \infty$), this is performed by the first multiplier, while at the bottom border ($-1 \leq u \leq +1, v = 0$), the equality is insured by the second operand.

This mapping may be used to solve several electrostatics problems with the geometry shown in Fig. 9; probably the most surprising of them is the following one. A straight gap of width $2t$ is cut in a thin conducting plane, and voltage V is applied between the resulting half-planes – see the bold lines in Fig. 10. Selecting a Cartesian coordinate system with axis z along the cut, axis y perpendicular to the plane, and the origin in the middle of the cut, we can write the boundary conditions of this Laplace problem as

$$\phi = \begin{cases} +V/2, & \text{at } x > t, y = 0, \\ -V/2, & \text{at } x < -t, y = 0. \end{cases} \quad (2.82)$$

(Due to problem's symmetry, we may expect that in the middle of the gap, i.e. at $-t < x < +t$ and $y = 0$, the electric field is parallel to the plane and hence $\partial\phi/\partial y = 0$.) The comparison of Figs. 9 and 10 shows that if we normalize our coordinates to t , Eq. (81) provides the conformal mapping of our system on plane z to the field in a plane capacitor on plane w , with voltage V between two planes $u = \pm 1$. Since we

³¹ See, e.g., MA Eq. (3.5).

already know that in that case $\phi = (V/2)u$, we may immediately use the first of Eqs. (81b) to write the final solution of the problem (in the dimensional coordinates):³²

$$\phi = \frac{V}{2}u = \frac{V}{\pi} \arcsin \frac{2x}{[(x+t)^2 + y^2]^{1/2} + [(x-t)^2 + y^2]^{1/2}}. \quad (2.83)$$

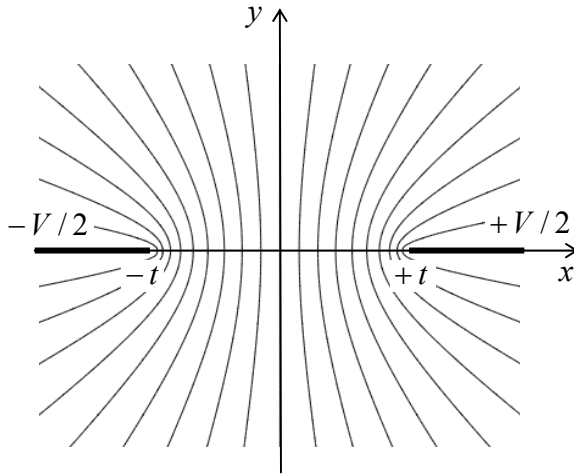


Fig. 2.10. Equipotential surfaces of the electric field between two thin conducting semi-planes (or rather their cross-sections by the perpendicular plane $z = \text{const}$).

Thin lines in Fig. 10 show the corresponding equipotential surfaces;³³ it is evident that the electric field concentrates at the gap edges, just as it did at the edge of the thin disk (Fig. 6). Let me leave the remaining calculation of the surface charge distribution and the mutual capacitance between the half-planes (per unit length) for reader's exercise.

2.5. Variable separation

The general approach of the methods discussed in the last two sections was to satisfy the Laplace equation by a function of a single variable that also satisfies the boundary conditions. Unfortunately, in many cases this cannot be done (at least, using practicably simple functions). In this case, a very powerful method, called *variable separation*, may work, frequently producing “semi-analytical” results in the form of an infinite series of either elementary or well-studied special functions. The main idea of the method is to present the solution of the general boundary problem (35) as the sum of partial solutions,

$$\phi = \sum_k c_k \phi_k, \quad (2.84)$$

where each function ϕ_k satisfies the Laplace equation, and then select the set of coefficients c_k to satisfy the boundary conditions. More specifically, in the variable separation method the partial solutions ϕ_k are looked for in the form of a product of functions, each depending of just one spatial coordinate.

³² This result could also be obtained using the so-called *elliptical* (not ellipsoidal!) coordinates.

³³ Another graphical representation of the electric field distribution, by *field lines*, is much less convenient. As a reminder, the field lines are defined as lines to whom the (in our current case, electrostatic) field vectors are tangential at each point. By this definition, the field lines are always normal to the equipotential surfaces, so that it is always straightforward to sketch them from the equipotential surface pattern – such as shown in Fig. 10.

(i) Cartesian coordinates. Let us discuss this approach on the classical example of a rectangular box with conducting walls (Fig. 11), with the same potential (that I will take for zero) at all the walls, but a different potential V fixed at the top lid. Moreover, in order to demonstrate the power of the variable separation method, let us carry out all the calculations for a more general case when the top lid potential is an arbitrary 2D function $V(x, y)$.³⁴

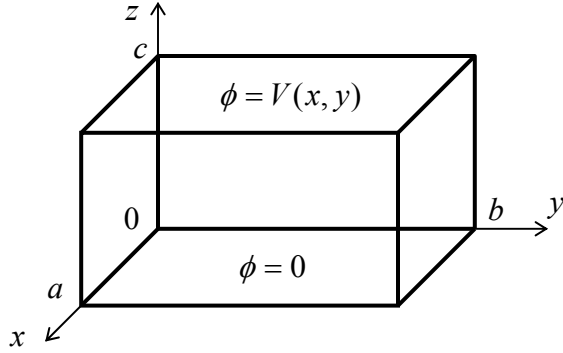


Fig. 2.11. Standard playground for the variable separation method discussion: a rectangular box with five conducting, grounded walls and a fixed potential distribution $V(x, y)$ on the top lid.

For this geometry, it is natural to use Cartesian coordinates $\{x, y, z\}$ and hence present each of the partial solutions in Eq. (84) as a product

$$\phi_k = X(x)Y(y)Z(z). \quad (2.85)$$

Plugging it into the Laplace equation expressed in the Cartesian coordinates,

$$\frac{\partial^2 \phi_k}{\partial x^2} + \frac{\partial^2 \phi_k}{\partial y^2} + \frac{\partial^2 \phi_k}{\partial z^2} = 0, \quad (2.86)$$

and dividing the result by product XYZ , we get

$$\frac{1}{X} \frac{d^2 X}{dx^2} + \frac{1}{Y} \frac{d^2 Y}{dy^2} + \frac{1}{Z} \frac{d^2 Z}{dz^2} = 0. \quad (2.87)$$

Here comes the punch line of the variable separation method: since the first term of this sum may depend only on x , the second one only of y , etc., Eq. (87) may be satisfied everywhere in the volume only if each of these terms equals a constant. In a minute we will see that for our current problem (Fig. 11), these constant x - and y -terms have to be negative; hence let us denote these *variable separation constants* as $(-\alpha^2)$ and $(-\beta^2)$, respectively. Now Eq. (87) shows that the constant z -term has to be positive; if we denote it as γ^2 , we get the following relation:

$$\alpha^2 + \beta^2 = \gamma^2. \quad (2.88)$$

Now the variables are separated in the sense that for functions $X(x)$, $Y(y)$, and $Z(z)$ we have got separate ordinary differential equations,

³⁴ Such distributions may be implemented in practice using so-called *mosaic electrodes* consisting of many electrically-insulated and individually-biased panels.

$$\frac{d^2 X}{dx^2} + \alpha^2 X = 0, \quad \frac{d^2 Y}{dy^2} + \beta^2 Y = 0, \quad \frac{d^2 Z}{dz^2} - \gamma^2 Z = 0, \quad (2.89)$$

which are related only by Eq. (88) for their parameters. Let us start from the equation for function $X(x)$. Its general solution is the sum of functions $\sin \alpha x$ and $\cos \alpha x$, multiplied by arbitrary coefficients. Let us select these coefficients to satisfy our boundary conditions. First, since $\phi \propto X$ should vanish at the back vertical wall of the box (i.e., with the choice of coordinate origin shown in Fig. 11, at $x = 0$ for any y and z), the coefficient at $\cos \alpha x$ should be zero. The remaining coefficient (at $\sin \alpha x$) may be included into the general factor c_k in Eq. (84), so that we may take X in the form

$$X = \sin \alpha x. \quad (2.90)$$

This solution satisfies the boundary condition at the opposite wall ($x = a$) only if its argument αa is a multiple of π , i.e. if α is equal to any of the following numbers (commonly called *eigenvalues*):³⁵

$$\alpha_n = \frac{\pi}{a} n, \quad n = 1, 2, \dots \quad (2.91)$$

(Terms with negative values of n would not be linearly-independent from those with positive n , and may be dropped from the sum (84). Value $n = 0$ is formally possible, but would give $X = 0$, i.e. $\phi_k = 0$, at any x , i.e. no contribution to sum (84), so it may be dropped as well.) Now we see that we indeed had to take α real, (i.e. α^2 positive); otherwise, instead of the oscillating function (90) we would have a sum of two exponential functions, which cannot equal zero in two independent points of axis x .

Since the equation for function $Y(y)$ is similar to that for $X(x)$, and the boundary conditions on the walls perpendicular to axis y ($y = 0$ and $y = b$) are similar to those for x -walls, the absolutely similar reasoning gives

$$Y = \sin \beta y, \quad \beta_m = \frac{\pi}{b} m, \quad m = 1, 2, \dots, \quad (2.92)$$

where the choice of integer m is independent of that of integer n . Now we see that according to Eq. (88), the separation constant γ depends on two indices, n and m , so that the relation may be rewritten as

$$\gamma_{nm} = [\alpha_n^2 + \beta_m^2]^{1/2} = \pi \left[\left(\frac{n}{a} \right)^2 + \left(\frac{m}{b} \right)^2 \right]^{1/2}. \quad (2.93)$$

The corresponding solution of the differential equation for Z may be presented as a sum of two exponents $\exp\{\pm \gamma_{nm} z\}$, or alternatively as a linear combination of two hyperbolic functions, $\sinh \gamma_{nm} z$ and $\cosh \gamma_{nm} z$, with arbitrary coefficients. At our choice of coordinate origin, the latter option is preferable, because $\cosh \gamma_{nm} z$ cannot satisfy the zero boundary condition at the bottom lid of the box ($z = 0$). Hence we may take Z in the form

$$Z = \sinh \gamma_{nm} z \quad (2.94)$$

³⁵ Note that according to Eqs. (91)-(92), as the spatial dimensions a and b of the system are increased, the distances between adjacent eigenvalues tend to zero. This fact implies that for spatially-infinite, non-periodic systems, the eigenvalue spectra are continuous, so that the sums of the type (84) become integrals. A few problems of this type are provided in Sec. 9 for reader's exercise.

that automatically satisfies that condition.

Now it is the right time to combine Eqs. (84) and (85) for our case in a more explicit form, replacing symbol k for the set of two integer indices n and m :

$$\phi(x, y, z) = \sum_{n,m=1}^{\infty} c_{nm} \sin \frac{\pi n x}{a} \sin \frac{\pi m y}{b} \sinh \gamma_{nm} z, \quad (2.95)$$

Variable
separation
in Cartesian
coordinates
(example)

where γ_{nm} is given by Eq. (93). This solution satisfies our boundary conditions on all walls of the box, besides the top lid, for arbitrary coefficients c_{nm} . The only job left for us is to choose these coefficients from the top-lid requirement:

$$\phi(x, y, c) = V(x, y) = \sum_{n,m=1}^{\infty} c_{nm} \sin \frac{\pi n x}{a} \sin \frac{\pi m y}{b} \sinh \gamma_{nm} c. \quad (2.96)$$

It seems like a bad luck to have just one equation for the infinite set of coefficients c_{nm} . However, the decisive help come from the fact that the functions of x and y that participate in Eq. (96), form *full, orthogonal* sets of 1D functions. The last term means that the integrals of the products of the functions with different integer indices over the region of interest equal zero. Indeed, direct integration gives

$$\int_0^a \sin \frac{\pi n x}{a} \sin \frac{\pi n' x}{a} dx = \begin{cases} a/2, & \text{for } n = n', \\ 0, & \text{for } n \neq n', \end{cases} \quad (2.97)$$

and similarly for y (with evident replacements $a \rightarrow b$, $n \rightarrow m$). Hence, the fruitful way to proceed is to multiply both sides of Eq. (96) by the product of the basis functions, with arbitrary indices n' and m' , and integrate the result over x and y :

$$\int_0^a dx \int_0^b dy V(x, y) \sin \frac{\pi n' x}{a} \sin \frac{\pi m' y}{b} = \sum_{n,m=1}^{\infty} c_{nm} \sinh \gamma_{nm} c \int_0^a \sin \frac{\pi n x}{a} \sin \frac{\pi n' x}{a} dx \times \int_0^b \sin \frac{\pi m y}{b} \sin \frac{\pi m' y}{b} dy. \quad (2.98)$$

Due to Eq. (97), all terms in the right-hand part of the last equation, besides those with $n = n'$ and $m = m'$, vanish, and (replacing n' with n , and m' with m) we finally get

$$c_{nm} = \frac{4}{ab \sinh \gamma_{nm} c} \int_0^a dx \int_0^b dy V(x, y) \sin \frac{\pi n x}{a} \sin \frac{\pi m y}{b}. \quad (2.99)$$

Relations (93), (95) and (99) present the complete solution of the posed boundary problem; we can see both good and bad news here. The first bit of bad news is that in the general case we still need to work out (formally, the infinite number of) integrals (99). In some cases, it is possible to do this analytically. For example, in our initial problem of constant potential on the top lid, $V(x, y) = \text{const} \equiv V_0$, both 1D integrations are elementary; for example

$$\int_0^a \sin \frac{\pi n x}{a} dx = \frac{2a}{\pi n} \times \begin{cases} 1, & \text{for } n \text{ odd,} \\ 0, & \text{for } n \text{ even,} \end{cases} \quad (2.100)$$

and similarly for the integral over y , so that

$$c_{nm} = \frac{16V_0}{\pi^2 nm \sinh \gamma_{nm} c} \times \begin{cases} 1, & \text{if both } n \text{ and } m \text{ are odd,} \\ 0, & \text{otherwise.} \end{cases} \quad (2.101)$$

The second bad news is that even at such a happy occasion, we still have to sum up the infinite series (95), so that our result may only be called analytical with some reservations, because in most cases we need a computer to get the final numbers or plots.

Now the first *good* news. Computers are very efficient for both operations (95) and (99), i.e. summation and integration. (As was discussed in Sec. 1.2, random errors are averaged out at these operations.) As an example, Fig. 12 shows the plots of the electrostatic potential in a cubic box ($a = b = c$), with an equipotential top lid ($V = V_0 = \text{const}$), obtained by numerical summation of series (95), using the analytical expression (101). The remarkable feature of this calculation is the very fast convergence of the series; for the middle cross-section of the cubic box ($z/c = 0.5$), already the first term (with $n = m = 1$) gives accuracy about 6%, while the sum of four leading terms (with $n, m = 1, 3$) reduces the error to just 0.2%. (For a longer box, $c > a, b$, the convergence is even faster – see the discussion below.) Only close to the corners between the top lid and the side walls, where the potential changes very rapidly, several more terms are necessary to get a reasonable accuracy.

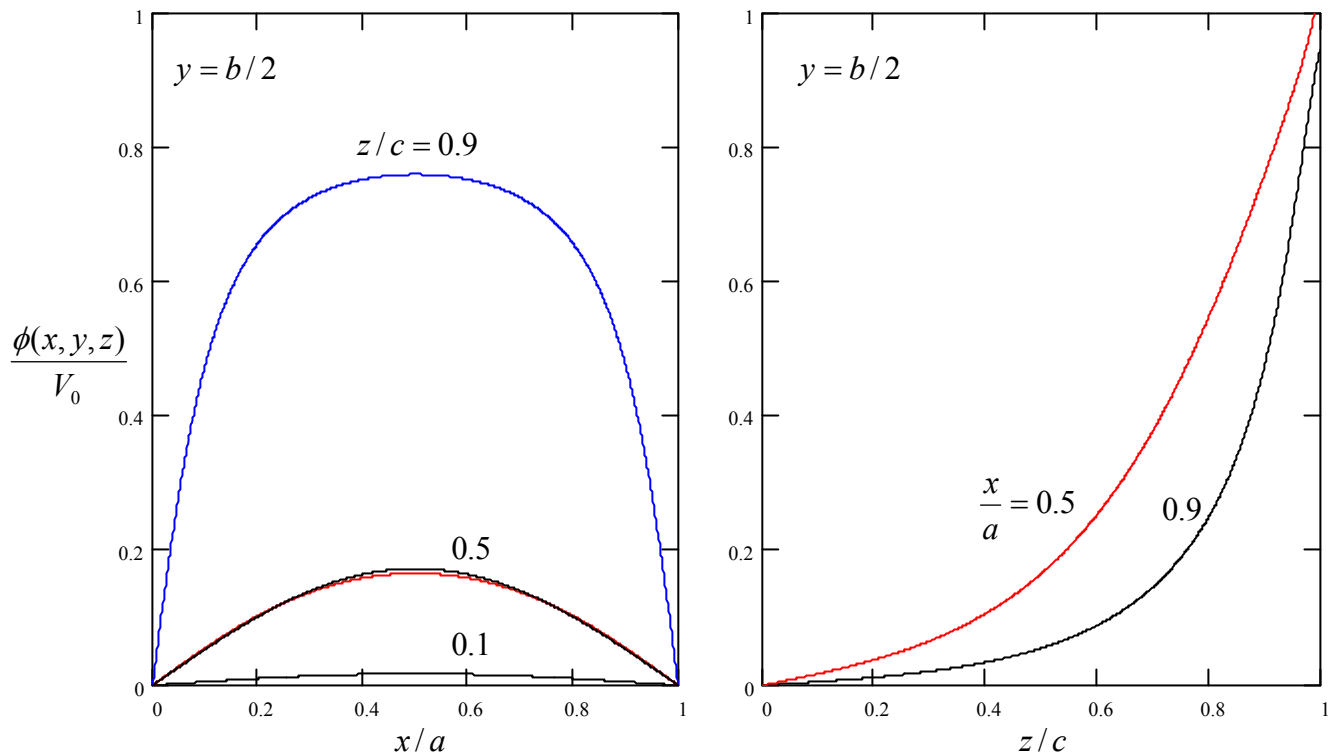


Fig. 2.12. Distribution of the electrostatic potential within a cubic box ($a = b = c$) with constant voltage V_0 on the top lid (Fig. 11), calculated numerically from Eqs. (93), (95) and (101). The dashed line on the left panel shows the contribution of the main term (with $n = m = 1$) to the full result.

The second good news is that our “semi-analytical” result allow its ultimate limits to be explored analytically. For example, Eq. (93) shows that for a very flat box ($c \ll a, b$), $\gamma_{n,m}z \leq \gamma_{n,m}c \ll 1$ at least for the lowest terms of series (95), with $n, m \ll c/a, c/b$. In these terms, sinh functions in Eqs. (96) and (99) may be well approximated with their arguments, and their ratio by z/c . This means that if we limit the summation to these term, Eq. (95) gives a very simple result

$$\phi(x, y) \approx \frac{z}{c} V(x, y) \quad (2.102)$$

which means that each segment of the flat box behaves just as a plane capacitor. Only near the vertical walls (or near possible locations where $V(x, y)$ is changed sharply), the higher terms in the series (95) are important, producing deviations from Eq. (102). In the opposite limit ($a, b \ll c$), Eq. (93) shows that, in contrast, $\gamma_{n,m}c \ll 1$ for all n and m . Moreover, the ratio $\sinh \gamma_{n,m}z / \sinh \gamma_{n,m}c$ drops sharply if either n or m is increased, if z is not too close to c . Hence in this case a very good approximation may be obtained by keeping just the leading term, with $n = m = 1$, in Eq. (95), so that the problem of summation disappears. (We saw above that this approximation works reasonably well even for a cubic box.) In particular, for the constant potential of the upper lid, we can use Eq. (101) and the exponential asymptotic for both \sinh functions, to get a very simple formula:

$$\phi = \frac{16}{\pi^2} \sin \frac{\pi x}{a} \sin \frac{\pi y}{b} \exp \left\{ -\pi \frac{(a^2 + b^2)^{1/2}}{ab} (c - z) \right\}. \quad (2.103)$$

The same variable separation method may be used to solve more general problems as well. For example, if all walls of the box shown in Fig. 11 have an arbitrary potential distribution, one can use the linear superposition principle to argue that the electrostatic potential distribution inside the box is the sum of 6 partial solutions of the type of Eq. (95), each with one wall biased by the corresponding voltage, and all other grounded ($\phi = 0$).

To summarize, the results given by the variable separation method are closer to what we could call a genuinely analytical solution than to purely numerical solutions - see Sec. 6 below. Now, let us explore the issues that arise when this method is applied in other orthogonal coordinate systems.

(ii) Polar coordinates. If a system of conductors is cylindrical, the potential distribution is independent of the coordinate z along the cylinder axis: $\partial\phi/\partial z = 0$, and the Laplace equation becomes two-dimensional. If conductor's cross-section is rectangular, the variable separation method works best in Cartesian coordinates $\{x, y\}$, and is just a particular case of the 3D solution discussed above. However, if the cross-section is circular, much more compact results may be obtained by using polar coordinates $\{\rho, \varphi\}$. As we already know from the last section, these 2D coordinates are orthogonal, so that the two-dimensional Laplace operator is a simple sum.³⁶ Requiring, just as we have done above, each component of sum (84) to satisfy the Laplace equation, we get

$$\frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho \frac{\partial \phi_k}{\partial \rho} \right) + \frac{1}{\rho^2} \frac{\partial^2 \phi_k}{\partial \varphi^2} = 0. \quad (2.104)$$

In a full analogy with Eq. (75), let us present each particular solution as a product: $\phi_k = \mathcal{R}(\rho) \mathcal{A}(\varphi)$. Plugging this expression into Eq. (104) and then dividing all its parts by $\mathcal{R}\mathcal{A}/\rho^2$, we get

$$\frac{\rho}{\mathcal{R}} \frac{d}{d\rho} \left(\rho \frac{d\mathcal{R}}{d\rho} \right) + \frac{1}{\mathcal{A}} \frac{d^2 \mathcal{A}}{d\varphi^2} = 0. \quad (2.105)$$

Following the same reasoning as for the Cartesian coordinates, we get two separated ordinary differential equations

³⁶ See, e.g., MA Eq. (10.3) with $\partial/\partial z = 0$.

$$\rho \frac{d}{d\rho} \left(\rho \frac{d\mathcal{R}}{d\rho} \right) = \nu^2 \mathcal{R}, \quad (2.106)$$

$$\frac{d^2 \mathcal{Z}}{d\varphi^2} + \nu^2 \mathcal{Z} = 0, \quad (2.107)$$

where ν^2 is the variable separation constant.

Let us start their analysis from Eq. (106), plugging into it a probe solution $\mathcal{R} = c\rho^\alpha$, where c and α are some constants. Elementary differentiation shows that if $\alpha \neq 0$, the equation is indeed satisfied for any c , with just one requirement on constant α , namely $\alpha^2 = \nu^2$. This means that the following linear superposition

$$\mathcal{R} = a_\nu \rho^{+\nu} + b_\nu \rho^{-\nu}, \quad \text{for } \nu \neq 0, \quad (2.108)$$

with constant coefficients a_ν and b_ν , is also a solution to Eq. (106). Moreover, the general theory of linear ordinary differential equations tells us that the solution of a second-order equation like Eq. (106) may only depend on just two constant factors that scale two linearly-independent functions. Hence, for all values $\nu^2 \neq 0$, Eq. (108) presents the *general* solution of that equation. The case when $\nu = 0$, in which functions $\rho^{+\nu}$ and $\rho^{-\nu}$ are just constants and hence are *not* linearly-independent, is special, but in this case the integration of Eq. (106) is straightforward,³⁷ giving

$$\mathcal{R} = a_0 + b_0 \ln \rho, \quad \text{for } \nu = 0. \quad (2.109)$$

In order to specify the separation constant, we should use Eq. (107), whose general solution is

$$\mathcal{Z} = \begin{cases} c_\nu \cos \nu\varphi + s_\nu \sin \nu\varphi, & \text{for } \nu \neq 0, \\ c_0 + s_0\varphi, & \text{for } \nu = 0. \end{cases} \quad (2.110)$$

There are two possible cases here. In many boundary problems solvable in cylindrical coordinates, the free space region, in which the Laplace equation is valid, extends continuously around the origin point $\rho = 0$. In this region, the potential has to be continuous and uniquely defined, so that \mathcal{Z} has to be a 2π -periodic function of angle φ . For that, one needs $\nu(\varphi + 2\pi)$ to be equal to $\nu\varphi + 2\pi n$, with n an integer, immediately giving us a discrete spectrum of possible values of the variable separation constant:

$$\nu = n = 0, \pm 1, \pm 2, \dots \quad (2.111)$$

In this case both functions \mathcal{R} and \mathcal{Z} may be labeled with the integer index n . Taking into account that the terms with negative values of n may be summed up with those with positive n , and that s_0 should equal zero (otherwise the 2π -periodicity of function \mathcal{Z} would be violated), we see that the general solution to the 2D Laplace equation may be presented as

$$\phi(\rho, \varphi) = a_0 + b_0 \ln \rho + \sum_{n=1}^{\infty} \left(a_n \rho^n + \frac{b_n}{\rho^n} \right) (c_n \cos n\varphi + s_n \sin n\varphi). \quad (2.112)$$

Let us see how all this machinery works on the classical problem of a round cylindrical conductor placed into an electric field that is uniform and perpendicular to cylinder's axis at large

³⁷ Actually, we have already done it in Sec. 3 – see Eq. (43).

distances - see Fig. 13a.³⁸ First of all, let us explore the effect of system's symmetries on coefficients in Eq. (112). Selecting the coordinate system as shown in Fig. 13a, and taking the cylinder's potential for zero, we immediately have $a_0 = 0$. Moreover, due to the mirror symmetry about plane $[x, z]$, the solution has to be an even function of angle φ , and hence all coefficients s_n should also equal zero. Also, at large distances ($\rho \gg R$) from the cylinder axis its effect on the electric field should vanish, and the potential should approach that of the uniform field $\mathbf{E} = E_0 \mathbf{n}_x$:

$$\phi \rightarrow -E_0 x = -E_0 \rho \cos \varphi, \quad \text{for } \rho \rightarrow \infty. \quad (2.113)$$

This is only possible if in Eq. (112), $b_0 = 0$, and also all coefficients a_n with $n \neq 1$ vanish, while product $a_1 c_1$ should be equal to $(-E_0)$. Thus the solution is reduced to the following form

$$\phi(\rho, \varphi) = -E_0 \rho \cos \varphi + \sum_{n=1}^{\infty} \frac{B_n}{\rho^n} \cos n\varphi, \quad (2.114)$$

in which coefficients $B_n \equiv b_n c_n$ should be found from the boundary condition on the cylinder's surface, i.e. at $\rho = R$:

$$\phi(R, \varphi) = 0. \quad (2.115)$$

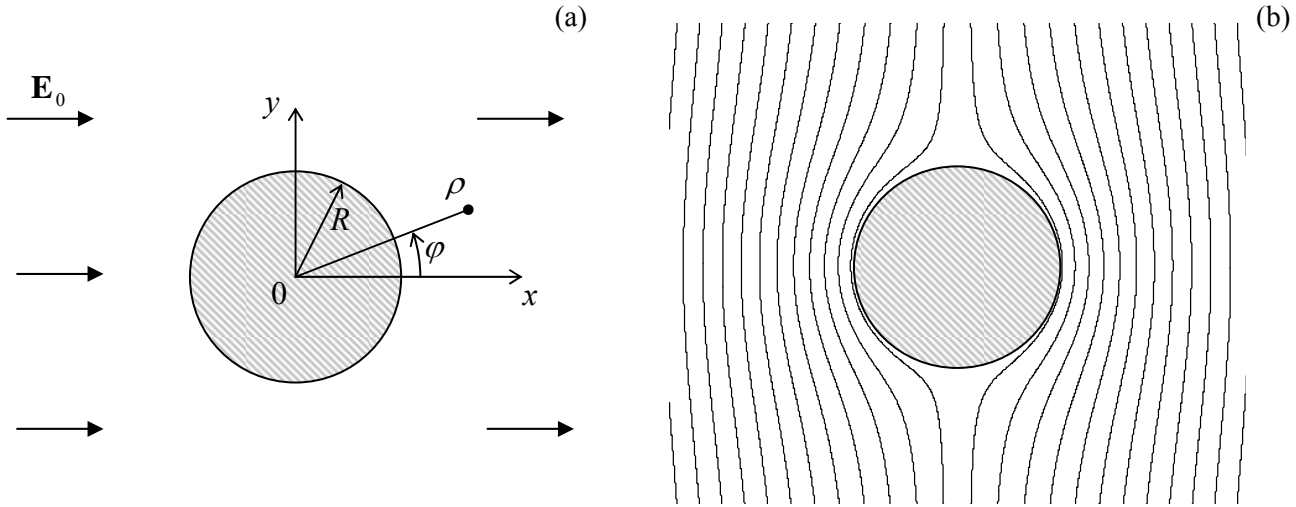


Fig. 2.13. Conducting cylinder inserted into an initially uniform electric field perpendicular to its axis: (a) the problem's geometry, and (b) the equipotential surfaces given by Eq. (117).

This requirement yields the following equation,

$$\left(\frac{B_1}{R} - E_0 R \right) \cos \varphi + \sum_{n=2}^{\infty} \frac{B_n}{R^n} \cos n\varphi = 0, \quad (2.116)$$

³⁸ This problem does belong to our current topic of electrostatic fields between conductors, because the uniform electric field may be created by a large plane capacitor.

which should be satisfied for all φ . But since functions $\cos n\varphi$ are orthogonal, this equality is only possible if all B_n for $n \geq 2$ are equal zero, while $B_1 = E_0 R^2$. Hence our final answer (which is of course only valid outside of the cylinder, i.e. for $\rho \geq R$), is

$$\phi(\rho, \varphi) = -E_0 \left(\rho - \frac{R^2}{\rho} \right) \cos \varphi = -E_0 \left(1 - \frac{R^2}{x^2 + y^2} \right) x. \quad (2.117)$$

This result (Fig. 13b) shows a smooth transition between the uniform field (113) far from the cylinder, to the equipotential surface of the cylinder (with $\phi = 0$). Such smoothening is very typical for Laplace equation solutions. Indeed, as we know from Chapter 1, these solutions corresponds to the lowest potential energy (1.67), and hence the lowest values of potential gradient modulus, possible at the given boundary conditions.

To complete the problem, let us calculate the distribution of the surface charge density over the cylinder's cross-section, using Eq. (3):

$$\sigma = \varepsilon_0 E_n|_{\text{surface}} = -\varepsilon_0 \frac{\partial \phi}{\partial \rho} \Big|_{\rho=R} = \varepsilon_0 E_0 \cos \varphi \frac{\partial}{\partial \rho} \left(\rho - \frac{R^2}{\rho} \right) \Big|_{\rho=R} = 2\varepsilon_0 E_0 \cos \varphi. \quad (2.118)$$

This very simple formula shows that at the field direction shown in Fig. 13a ($E_0 > 0$), the surface charge is positive on the right side of the cylinder and negative on its left side, thus creating a field directed from the right to the left, that compensates the external field inside the conductor, where the net field is zero. Note also that the net electric charge of the cylinder is zero, in the correspondence with the problem symmetry. Another useful by-product of calculation (118) is that the surface electric field equals $2E_0 \cos \varphi$, and hence its largest magnitude is twice the field far from the cylinder. Such electric field concentration is very typical for all convex conducting surfaces.

The last observation gets additional confirmation for the second possible topology, when Eq. (110) is used to describe problems with no angular periodicity. A typical example is a cylindrical conductor with a cross-section that features a corner limited by straight lines (Fig. 14). Indeed, at we may argue that at $\rho < R$ (where R is the scale of radial extension of the straight sides of the corner), the Laplace equation may be satisfied by a sum of partial solutions $\mathcal{A}(\rho)\mathcal{A}(\varphi)$ if the angular components of the products satisfy the boundary conditions on the corner sides. Taking (just for the simplicity of notation) the conductor's potential to be zero, and one of the corner's sides as axis x ($\varphi = 0$), these boundary conditions are

$$\mathcal{A}(0) = \mathcal{A}(\beta) = 0, \quad (2.119)$$

where angle β may be anywhere between 0 and 2π (Fig. 14).

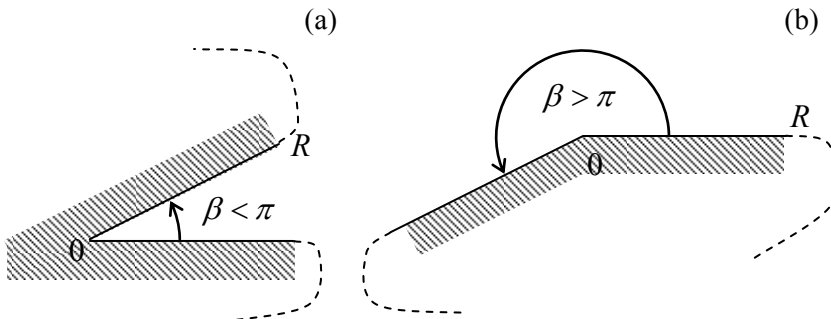


Fig. 2.14. Cylindrical conductor cross-sections with (a) a corner and (b) a wedge.

Comparing this condition with Eq. (110), we see that it requires c_ν to vanish, and ν to take one of the values of the following discrete spectrum:

$$\nu_m = (\pi / \beta)m, \quad (2.120)$$

with positive integer m . Hence the full solution of the Laplace equation takes the form

$$\phi = \sum_{m=1}^{\infty} a_m \rho^{\pi m / \beta} \sin \frac{\pi m \varphi}{\beta}, \quad \text{for } \rho < R, \quad (2.121)$$

where constants s_ν have been incorporated into a_m . The set of constants a_m cannot be simply determined, because it depends on the exact shape of the conductor outside the corner, and the externally applied electric field. However, whatever the set is, in the limit $\rho \rightarrow 0$, solution (121) is almost³⁹ always dominated by the term with lowest ν (corresponding to $m = 1$),

$$\phi \rightarrow a_1 \rho^{\pi / \beta} \sin \frac{\pi}{\beta} \varphi, \quad (2.122)$$

because the higher terms go to zero faster. This potential distribution corresponds to the surface charge density

$$\sigma = \varepsilon_0 E_n|_{\text{surface}} = -\varepsilon_0 \frac{\partial \phi}{\partial(\rho \varphi)} \Big|_{\rho=\text{const}, \varphi \rightarrow +0} = -\varepsilon_0 \frac{\pi a_1}{\beta} \rho^{(\pi / \beta - 1)}. \quad (2.123)$$

(It is similar on the opposite face of the angle.)

Equation (123) shows that if we are dealing with a usual, concave corner ($\beta < \pi$, see Fig. 14a), the charge density (and the surface electric field) tends to zero. On the other case, at a “convex corner” with $\beta > \pi$ (actually, a wedge - see Fig. 14b), both charge and field concentrate, formally diverging at $\rho \rightarrow 0$. (So, do not sit on a roof’s ridge during a thunderstorm; rather hide in a ditch!) We already saw qualitatively similar effects at our analyses of the thin round disk and split plane in the past section.

(iii) Cylindrical coordinates. Now, let us discuss whether it is possible to generalize our approach to problems whose geometry is still axially-symmetric, but with a substantial dependence of the potential on the axial coordinate ($\partial \phi / \partial z \neq 0$). The classical example of such a problem is shown in Fig. 15. Here the side wall and the bottom lid of a round cylinder are kept at fixed potential (say, $\phi = 0$), but the potential V fixed at the top lid is different. This problem is qualitatively similar to the rectangular box problem solved above (Fig. 11), and we will also try to solve it for the case of arbitrary voltage distribution over the top lid: $V = V(\rho, \varphi)$.

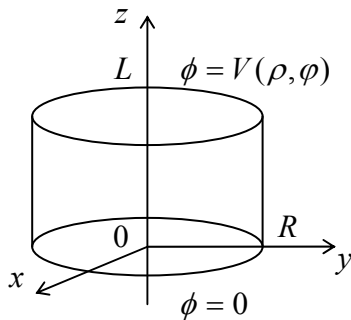


Fig. 2.15. Round cylinder with conducting walls.

³⁹ Exceptions are possible only for highly symmetric configurations when the external field are crafted to make $a_1 = 0$. In this case the solution is led by the first nonvanishing term of the series (121).

Following the main idea of the variable separation method, let us require that each partial function ϕ_k in Eq. (84) satisfies the Laplace equation, now in full cylindrical coordinates $\{\rho, \varphi, z\}$:⁴⁰

$$\frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho \frac{\partial \phi_k}{\partial \rho} \right) + \frac{1}{\rho^2} \frac{\partial^2 \phi_k}{\partial \varphi^2} + \frac{\partial^2 \phi_k}{\partial z^2} = 0. \quad (2.124)$$

Plugging in ϕ_k in the form $\mathcal{R}(\rho)\mathcal{A}(\varphi)\mathcal{Z}(z)$ into Eq. (124) and dividing both parts by product $\mathcal{R}\mathcal{A}\mathcal{Z}$, we get

$$\frac{1}{\rho\mathcal{R}} \frac{d}{d\rho} \left(\rho \frac{d\mathcal{R}}{d\rho} \right) + \frac{1}{\rho^2\mathcal{A}} \frac{d^2\mathcal{A}}{d\varphi^2} + \frac{1}{\mathcal{Z}} \frac{d^2\mathcal{Z}}{dz^2} = 0. \quad (2.125)$$

Since the first two terms of Eq. (125) can only depend on polar variables ρ and φ , while the third term, only on z , at least that term should be a constant. Denoting it (just like in the rectangular box problem) by γ^2 , we get, instead of Eq. (125), a set of two equations:

$$\frac{d^2\mathcal{Z}}{dz^2} = \gamma^2 \mathcal{Z}, \quad (2.126)$$

$$\frac{1}{\rho\mathcal{R}} \frac{d}{d\rho} \left(\rho \frac{d\mathcal{R}}{d\rho} \right) + \gamma^2 + \frac{1}{\rho^2\mathcal{A}} \frac{d^2\mathcal{A}}{d\varphi^2} = 0. \quad (2.127)$$

Now, multiplying all the terms of Eq. (127) by ρ^2 , we see that the last term, $(d^2\mathcal{A}/d\varphi^2)/\mathcal{A}$, may depend only on φ , and thus should be constant. Calling that constant ν^2 (as in Sec. (ii) above), we separate Eq. (127) into an angular equation,

$$\frac{d^2\mathcal{A}}{d\varphi^2} + \nu^2 \mathcal{A} = 0, \quad (2.128)$$

and a radial equation:

$$\frac{d^2\mathcal{R}}{d\rho^2} + \frac{1}{\rho} \frac{d\mathcal{R}}{d\rho} + \left(\gamma^2 - \frac{\nu^2}{\rho^2} \right) \mathcal{R} = 0. \quad (2.129)$$

We see that the ordinary differential equations for functions $\mathcal{Z}(z)$ and $\mathcal{A}(\varphi)$ (and hence their solutions) are identical to those discussed earlier in this section. However, Eq. (129) for the radial function $\mathcal{R}(\rho)$ (called the *Bessel equation*) is more complex than in the 2D case, and depends on two independent constant parameters, γ and ν . The latter challenge may be readily overcome if we notice that any change of γ may be reduced to re-scaling the radial coordinate ρ . Indeed, introducing a dimensionless variable $\xi \equiv \gamma\rho$,⁴¹ Eq. (129) may be reduced to an with one parameter, ν .

Bessel
equation

$$\frac{d^2\mathcal{R}}{d\xi^2} + \frac{1}{\xi} \frac{d\mathcal{R}}{d\xi} + \left(1 - \frac{\nu^2}{\xi^2} \right) \mathcal{R} = 0. \quad (2.130)$$

⁴⁰ See, e.g., MA Eq. (10.3).

⁴¹ Please note that this normalization is specific for each value of the variable separation parameter γ . Also, note that the normalization is meaningless for $\gamma = 0$, i.e. for the case $\mathcal{Z}(z) = \text{const}$. However, if we need partial solutions for this value of γ , we can use Eqs. (108)-(109).

Moreover, we already know that for angle-periodic problems the spectrum of eigenvalues of Eq. (128) is discrete $\nu = n$.

Unfortunately, even in this case, Eq. (130) cannot be satisfied by a single “elementary” function, and is the canonical form of an equation defining the *Bessel function of the first kind, of order ν* , commonly denoted as $J_\nu(\xi)$. Let me review in brief the Bessel function properties most relevant for the boundary problems of physics - and some other problems discussed in these notes.⁴²

First of all, the Bessel function of a negative integer order is very simply related to that with the positive order:

$$J_{-n}(\xi) = (-1)^n J_n(\xi), \quad (2.131)$$

enabling us to limit our discussion to the functions with $n \geq 0$. Figure 16 shows four functions with a few lowest positive n .

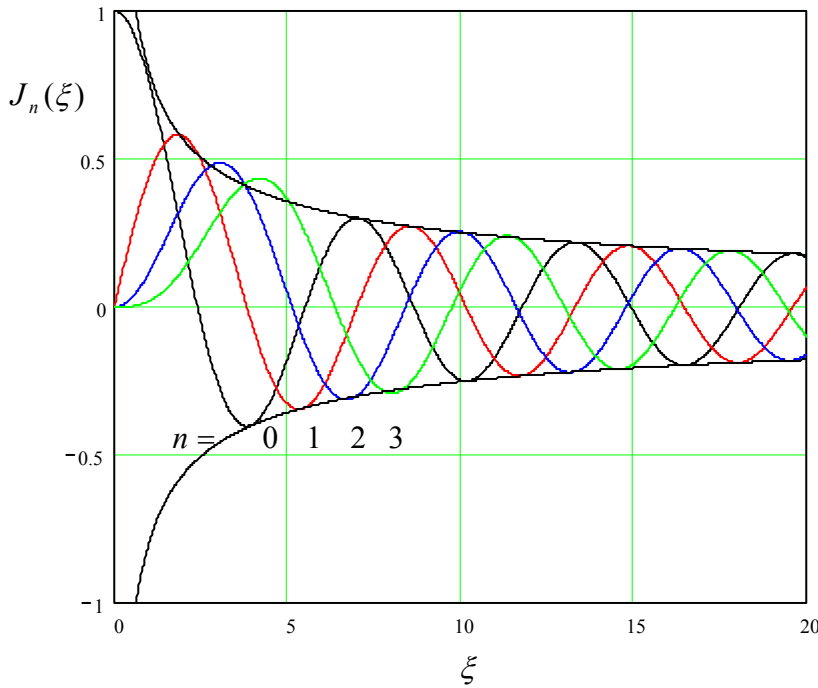


Fig. 2.16. Several first-kind Bessel functions $J_n(\xi)$ of integer order. Dashed lines show the envelope of asymptotes (135).

As argument x is increased, each function is initially close to a power law: $J_0(\xi) \approx 1$, $J_1(\xi) \approx \xi/2$, $J_2(\xi) \approx \xi^2/8$, etc. This behavior follows from the Taylor series

$$J_n(\xi) = \left(\frac{\xi}{2}\right)^n \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(n+k)!} \left(\frac{\xi}{2}\right)^{2k}, \quad (2.132)$$

which that is formally valid for any ξ , and may even serve as an alternative definition of function $J_n(\xi)$. However, this series is converging fast only at relatively small arguments, $\xi < n$, where its main term is

⁴² For a more complete discussion of these functions, see the literature listed in MA Sec. 16, for example, Chapter 6 (written by P. Davis) in the collection compiled and edited by Abramowitz and Stegun.

$$J_n(\xi) \Big|_{\xi \rightarrow 0} \rightarrow \frac{1}{n!} \left(\frac{\xi}{2} \right)^n. \quad (2.133)$$

At $\xi \approx n + 1.86n^{1/3}$, the Bessel function reaches its maximum⁴³

$$\max_{\xi} [J_n(\xi)] \approx \frac{0.675}{n^{1/3}}, \quad (2.134)$$

and then starts to oscillate with a period that gradually approaches 2π , a phase shift that increases by $\pi/2$ with each unit increment of n , and an amplitude that decreases as $\xi^{1/2}$. These features are described by the following asymptotic formula

$$J_n(\xi) \rightarrow \left(\frac{2}{\pi\xi} \right)^{1/2} \cos\left(\xi - \frac{\pi}{4} - \frac{n\pi}{2}\right), \quad \text{for } \xi/n \rightarrow \infty, \quad (2.135)$$

that starts to give reasonable results very soon above the function peaks – see Fig. 16.⁴⁴

Now we are ready to return to our case study (Fig. 15). Let us select functions $Z(z)$ to satisfy the bottom-lid boundary condition $Z(0) = 0$, i.e. proportional to $\sinh \gamma z$ – cf. Eq. (95). Then

$$\phi = \sum_{n=0}^{\infty} \sum_{\gamma} J_n(\gamma \rho) (c_{n\gamma} \cos n\varphi + s_{n\gamma} \sin n\varphi) \sinh \gamma z. \quad (2.136)$$

Next, we need to satisfy the zero boundary condition at the cylinder's side wall ($\rho = R$). This may be ensured by taking

$$J_n(\gamma R) = 0. \quad (2.137)$$

Since each function $J_n(x)$ has an infinite number of positive zeros (see Fig. 16), which may be numbered by an integer index $m = 1, 2, \dots$, Eq. (137) may be satisfied with an infinite number of discrete values of the separation parameter γ .

$$\gamma_{nm} = \frac{\xi_{nm}}{R}, \quad (2.138)$$

where ξ_{nm} is the m -th zero of function $J_n(x)$ – see the top numbers in the cells of Table 1. (Very soon we will see what do we need the bottom numbers for.)

Hence, Eq. (136) may be presented in a more explicit form:

$$\phi(\rho, \varphi, z) = \sum_{n=0}^{\infty} \sum_{m=1}^{\infty} J_n\left(\xi_{nm} \frac{\rho}{R}\right) (c_{nm} \cos n\varphi + s_{nm} \sin n\varphi) \sinh\left(\xi_{nm} \frac{z}{R}\right). \quad (2.139)$$

Variable
separation in
cylindrical
coordinates
(example)

⁴³ These two formulas for the Bessel function peak are strictly valid for $n \gg 1$, but may be used for reasonable estimates starting already from $n = 1$; for example, $\max_{\xi} [J_1(\xi)]$ is close to 0.58 and is reached at $\xi \approx 2.4$, just about 30% away from the values given by the asymptotic formulas.

⁴⁴ Eq. (135) and Fig. 16 clearly show the close analogy between the Bessel functions and the usual trigonometric functions, sine and cosine. In order to emphasize this similarity, and help the reader to develop more gut feeling of the Bessel functions, let me mention one fact of the elasticity theory: while sine functions describe, in particular, possible modes of standing waves on a guitar string, functions $J_n(\xi)$ describe, in particular, possible standing waves on an elastic round membrane, with $J_0(\xi)$ describing their lowest (fundamental) mode.

Here coefficients c_{nm} and s_{nm} have to be selected to satisfy the only remaining boundary condition – that on the top lid:

$$V(\rho, \varphi) = \phi(\rho, \varphi, L) = \sum_{n=0}^{\infty} \sum_{m=1}^{\infty} J_n(\xi_{nm} \frac{\rho}{R}) (c_{nm} \cos n\varphi + s_{nm} \sin n\varphi) \sinh\left(\xi_{nm} \frac{L}{R}\right). \quad (2.140)$$

To use it, let us multiply both parts of Eq. (140) by $J_n(\xi_{nm} \rho/R) \cos n'\varphi$, integrate the result over the lid area, and use the following property of the Bessel functions:

$$\int_0^1 J_n(\xi_{nm} s) J_n(\xi_{nm'} s) s ds = \frac{1}{2} [J_{n+1}(\xi_{nm})]^2 \delta_{nm'}, \quad (2.141)$$

where $\delta_{nm'}$ is the Kronecker symbol.⁴⁵

Table 2.1. Approximate values of a few first zeros of a few lowest-order Bessel functions $J_n(\xi)$ (the top number in each cell), and the values of $dJ_n/d\xi$ at those points (the bottom number in the cell).

	$m = 1$	2	3	4	5	6
$n = 0$	2.40482 -0.51914	5.52008 +0.34026	8.65372 -0.27145	11.79215 +0.23245	14.93091 -0.20654	18.07106 +0.18773
1	3.83171 -0.40276	7.01559 +0.30012	10.17347 -0.24970	13.32369 +0.21836	16.47063 -0.19647	19.61586 +0.18006
2	5.13562 -0.33967	8.41724 +0.27138	11.61984 -0.23244	14.79595 +0.20654	17.95982 -0.18773	21.11700 +0.17326
3	6.38016 -0.29827	9.76102 +0.24942	13.01520 -0.21828	16.22347 +0.19644	19.40942 -0.18005	22.58273 +0.16718
4	7.58834 -0.26836	11.06471 +0.23188	14.37254 -0.20636	17.61597 +0.18766	20.82693 -0.17323	24.01902 +0.16168
5	8.77148 -0.24543	12.33860 +0.21743	15.70017 -0.19615	18.98013 +0.17993	22.21780 -0.16712	25.43034 +0.15669

Relation (141) expresses a very specific (“2D”) orthogonality of Bessel functions with different indices m - do not confuse them with the function’s order n , please!⁴⁶ Since it relates two Bessel functions with the same index n , it is natural to ask why its right-hand part contains the function with a different index ($n + 1$). Some clue may come from one more very important property of the Bessel functions, the so-called *recurrence relations*:⁴⁷

⁴⁵ Let me hope the reader knows what it is; if not – see MA Eq. (13.1).

⁴⁶ The Bessel functions of the *same argument* but of *different orders* are also orthogonal, but in a different way:

$$\int_0^{\infty} J_n(\xi) J_{n'}(\xi) \frac{d\xi}{\xi} = \frac{1}{n + n'} \delta_{nn'}.$$

⁴⁷ These relations provide, in particular, a convenient way for fast numerical computation of all $J_n(\xi)$ after $J_0(\xi)$ has been computed. (The latter is usually done with an algorithm using Eq. (132) for smaller ξ and an extension of Eq. (135) for larger ξ .) Note that most mathematical software packages, including all those listed in MA Sec. 16(iv), include ready subroutines for calculation of functions $J_n(\xi)$ and other special functions used in this lecture series. In this sense, the line separated these “special functions” from “elementary functions” is rather blurry.

$$J_{n-1}(\xi) + J_{n+1}(\xi) = \frac{2nJ_n(\xi)}{\xi}, \quad (2.142a)$$

$$J_{n-1}(\xi) - J_{n+1}(\xi) = 2 \frac{dJ_n(\xi)}{d\xi}, \quad (2.142b)$$

that in particular yield the following relation (convenient for working out some Bessel function integrals):

$$\frac{d}{d\xi}(\xi^n J_n(\xi)) = \xi^n J_{n-1}(\xi). \quad (2.143)$$

For our current purposes, let us apply the recurrence relations at special points ξ_{nm} . At these points, J_n vanishes, and the system of two equations (142) may be readily solved to get, in particular,

$$J_{n+1}(\xi_{nm}) = -\frac{dJ_n}{d\xi}(\xi_{nm}), \quad (2.144)$$

so that the square bracket in the right-hand part of Eq. (141) is just $(dJ_n/d\xi)^2$ at $\xi = \xi_{nm}$. Thus the values of the Bessel function derivatives at the zero points (given by the lower numbers in the cells of Table 1) are as important for boundary problem solutions as the zeros themselves.

Since the angular functions $\cos n\varphi$ are also orthogonal – both to each other,

$$\int_0^{2\pi} \cos(n\varphi) \cos(n'\varphi) d\varphi = \pi \delta_{nn'}, \quad (2.145)$$

and to all functions $\sin n\varphi$, the integration over the lid area kills all terms of both series in right-hand part of Eq. (140), besides just one term proportional to $c_{n'm'}$, and hence gives an explicit expression for that coefficient. The counterpart coefficients $s_{n'm'}$ may be found by repeating the same procedure with the replacement of $\cos n'\varphi$ by $\sin n'\varphi$. This evaluation (left for reader's exercise) completes the solution of our problem for an arbitrary lid potential $V(\rho, \varphi)$.

Still, before leaving the Bessel functions (for a while :-), we need to address two important issues. First, we have seen that in our cylinder problem (Fig. 15), the set of functions $J_n(\xi_{nm}\rho/R)$ with different indices m (that characterize the degree of Bessel function's stretch along axis ρ) play the role similar to that of functions $\sin(\pi nx/a)$ in the rectangular box problem shown in Fig. 11. In this context, what is the analog of functions $\cos(\pi nx/a)$ – which may be important for some boundary problems? In a more formal language, are there any functions of the same argument $\xi \equiv \xi_{nm}\rho/R$, that would be linearly independent of the Bessel functions of the first kind, while satisfying the same differential equation (130)?

The answer is *yes*. For the definition of such functions, we first need to generalize our prior formulas for $J_n(\xi)$, and in particular Eq. (132), to the case of arbitrary order ν . The generalization may be performed in the following way:

$$J_\nu(\xi) = \left(\frac{\xi}{2}\right)^\nu \sum_{k=0}^{\infty} \frac{(-1)^k}{k! \Gamma(\nu + k + 1)} \left(\frac{\xi}{2}\right)^{2k}, \quad (2.146)$$

where $\Gamma(s)$ is the so-called *gamma function* that may be defined, for almost any real s , as⁴⁸

$$\Gamma(s) \equiv \int_0^{\infty} \xi^{s-1} e^{-\xi} d\xi. \quad (2.147)$$

The simplest, and the most important property of the gamma function is that for integer values of argument it gives the factorial of a number smaller by one:

$$\Gamma(n+1) = n! \equiv 1 \cdot 2 \cdot \dots n, \quad (2.148)$$

so it is essentially a generalization of the notion of factorial to all real numbers.

The Bessel functions defined by Eq. (146) satisfy (after replacements $n \rightarrow \nu$ and $n! \rightarrow \Gamma(n+1)$), virtually all the relations we have discussed above, including the Bessel equation (130), the asymptotic formula (135), the orthogonality condition (141), and the recurrence relations (142). Moreover, it may be shown that $\nu \neq n$, functions $J_\nu(\xi)$ and $J_{-\nu}(\xi)$ are linearly independent and hence their linear combination may be used to present a general solution of the Bessel equation. Unfortunately, as Eq. (131) shows, for $\nu = n$ this is not true, and a solution independent of $J_n(\xi)$ has to be formed in a different way.

The most common way of overcoming this difficulty is first to define, for all $\nu \neq n$, function

$$Y_\nu(\xi) \equiv \frac{J_\nu(\xi) \cos \nu\pi - J_{-\nu}(\xi)}{\sin \nu\pi}, \quad (2.149)$$

called the *Bessel function of second kind*, or more often as the *Weber functions*,⁴⁹ and then to follow the limit $\nu \rightarrow n$. At this, both the nominator and denominator of the right-hand part of Eq. (149) tend to zero, but their ratio tends to a finite value called $Y_n(x)$. It may be shown that these functions are still the solutions of the Bessel equation and are linearly independent of $J_n(x)$, though are related just as those functions if the sign of n changes:

$$Y_{-n}(\xi) = (-1)^n Y_n(\xi). \quad (2.150)$$

Figure 17 shows a few Weber functions of the lowest integer orders. The plots show that the asymptotic behavior is very much similar to that of $J_n(\xi)$,

$$Y_n(\xi) \rightarrow \left(\frac{2}{\pi\xi} \right)^{1/2} \sin\left(\xi - \frac{\pi}{4} - \frac{n\pi}{2}\right), \quad \text{for } \xi \rightarrow \infty, \quad (2.151)$$

but with the phase shift necessary to make these Bessel functions orthogonal to those of the first order – cf. Eq. (135). However, for small values of argument ξ , the Bessel functions of the second kind behave completely differently from those of the first kind:

$$Y_n(\xi) \rightarrow \begin{cases} (2/\pi) [\ln(\xi/2) + \gamma], & \text{for } n = 0, \\ -\frac{(n-1)!}{\pi} \left(\frac{\xi}{2} \right)^{-n}, & \text{for } n \neq 0, \end{cases} \quad (2.152)$$

⁴⁸ See, e.g., MA Eq. (6.7a). I used word “almost” because the gamma-function tends to infinity at all non-positive integer values of its argument ($s = 0, -1, -2, \dots$).

⁴⁹ They are also sometimes called the *Neumann functions*, and denoted as $N_\nu(\xi)$.

where

$$\gamma \equiv \lim_{n \rightarrow \infty} \left(1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} - \ln n \right) \approx 0.577157 \dots \quad (2.153)$$

is the so-called *Euler constant*. Relations (152) and Fig. 17 show that functions $Y_n(\xi)$ diverge at $\xi \rightarrow 0$ and hence cannot describe the behavior of any physical variable, in particular the electrostatic potential.

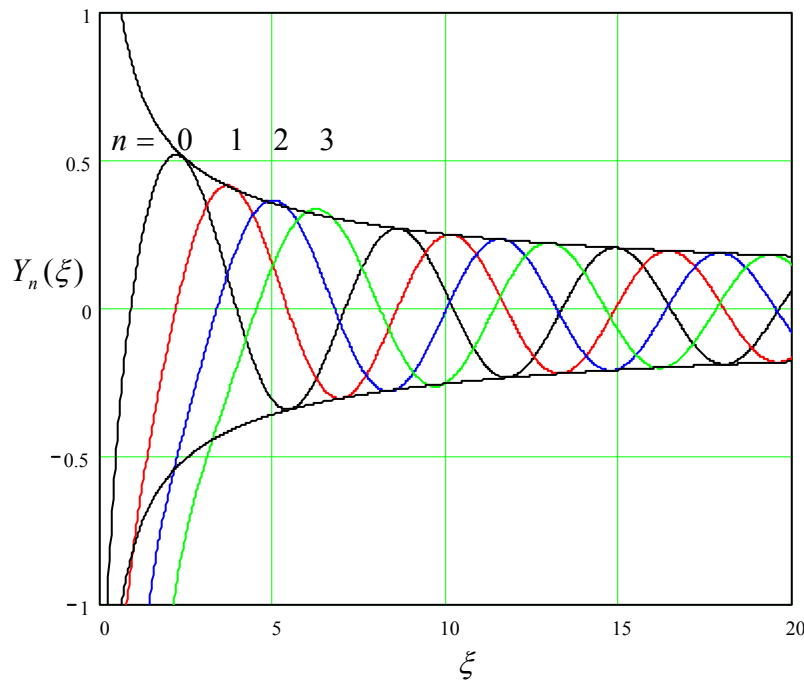


Fig. 2.17. A few Bessel functions of the second kind (a.k.a. the Neumann functions, a.k.a. the Weber functions).

One may wonder: if this is true, when do we need these functions in physics? This does not happen too often, but still does. Figure 18 shows an example of a boundary problem of electrostatics that requires both functions $J_n(\xi)$ and $Y_n(\xi)$.

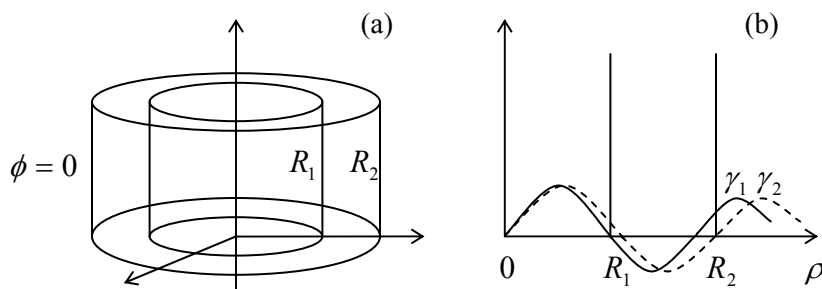


Fig. 2.18. Simple boundary problem that cannot be solved using just one kind of Bessel functions.

Two round, coaxial conducting cylinders are kept at the same (say, zero) potential, but at least one of two horizontal lids has a different potential. The problem is almost completely similar to that discussed above (Fig. 15), but now we need to find the potential distribution in the free space between the cylinders, $R_1 < \rho < R_2$. If we use the same variable separation as in the simpler counterpart problem,

we need the radial functions $\mathcal{R}(\rho)$ to satisfy two zero boundary conditions: at $\rho = R_1$ and $\rho = R_2$. With the Bessel functions of just first kind, $J_n(\gamma\rho)$, it is impossible to do, because the two boundaries would impose two independent (and generally incompatible) conditions, $J_n(\gamma R_1) = 0$, and $J_n(\gamma R_2) = 0$, for one “compression parameter” γ . The existence of the Bessel functions of the second kind immediately saves the day, because if a solution is presented as a linear combination,⁵⁰

$$c_J J_n(\gamma\rho) + c_Y Y_n(\gamma\rho), \quad (2.154)$$

two zero boundary conditions give two equations for γ and ratio $c \equiv c_Y/c_J$. (Due to the oscillating character of both Bessel functions, these conditions would be typically satisfied by an infinite set of discrete pairs $\{\gamma, c\}$.) Note, however, that generally none of these pairs would correspond to zeros of either J_n nor Y_n , so that having an analog of Table 1 for the latter function would not help much. Hence, even the simple problems of this kind (like the one shown in Fig. 18) typically require numerical solutions of algebraic (transcendental) equations.

One more issue we need to address, before moving on to the spherical coordinates, are the so-called *modified Bessel functions*: of the *first kind*, $I_\nu(\xi)$, and of the *second kind*, $K_\nu(\xi)$. They are two linearly-independent solutions of the *modified Bessel equation*,

$$\frac{d^2 \mathcal{R}}{d\xi^2} + \frac{1}{\xi} \frac{d\mathcal{R}}{d\xi} - \left(1 + \frac{\nu^2}{\xi^2}\right) \mathcal{R} = 0, \quad (2.155)$$

Modified
Bessel
equation

that differs from Eq. (130) “only” by the sign of one of its terms. Figure 19 shows a simple problem that leads to this equation: a round conducting cylinder is sliced, perpendicular to its axis, to rings of equal height h , which are kept at equal but sign-alternating potentials.

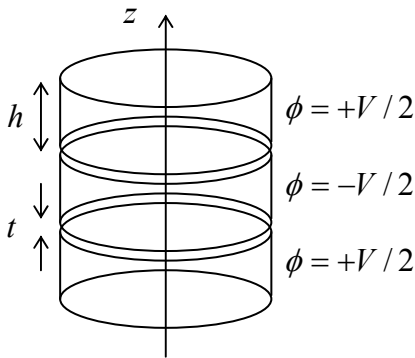


Fig. 2.19. Typical boundary problem whose solution may be conveniently described in terms of the modified Bessel functions.

If the gaps between the sections are narrow, $t \ll h$, we may use the variable separation method for the solution to this problem, but now we evidently need periodic (rather than exponential) solutions

⁵⁰ A pair of independent linear functions, used for presentation of the general solution of the Bessel equation, may be also chosen in a different way, using the so-called *Hankel functions*

$$H_n^{(1,2)}(\xi) \equiv J_n(\xi) \pm iY_n(\xi).$$

For representing the general solution of Eq. (130), this alternative is completely similar to using the pair of complex functions $\exp\{\pm i\alpha x\} = \cos \alpha x \pm i \sin \alpha x$ instead of the pair of real functions $\{\cos \alpha x, \sin \alpha x\}$ for representing the general solution of Eq. (89) for $X(x)$.

along axis z , i.e. linear combinations of $\sinh kz$ and $\cosh kz$ with various real values of constant k . Separating the variables, we arrive at a differential equation similar to Eq. (129), but with the negative sign before the separation constant:

$$\frac{d^2 \mathcal{R}}{d\rho^2} + \frac{1}{\rho} \frac{d\mathcal{R}}{d\rho} - (k^2 + \frac{\nu^2}{\rho^2}) \mathcal{R} = 0. \quad (2.156)$$

Radial coordinate normalization, $\xi \equiv k\rho$, immediately leads us to Eq. (155), and hence (for $\nu = n$) to the modified Bessel functions $I_n(\xi)$ and $K_n(\xi)$.

Figure 19 shows the behavior of a few such functions, of a few lowest orders. One can see that at $\xi \rightarrow 0$ it is virtually similar to that of the “usual” Bessel functions - cf. Eqs. (132) and (152), with $K_n(\xi)$ multiplied (due to purely historical reasons) by an additional coefficient, $\pi/2$:

$$I_n(\xi) \rightarrow \frac{1}{n!} \left(\frac{\xi}{2} \right)^n, \quad K_n(\xi) \rightarrow \begin{cases} -\left[\ln\left(\frac{\xi}{2}\right) + \gamma \right], & \text{for } n = 0, \\ \frac{(n-1)!}{2} \left(\frac{\xi}{2} \right)^{-n}, & \text{for } n \neq 0, \end{cases} \quad (2.157)$$

However, the asymptotic behavior of the modified functions is very much different, with $I_n(x)$ exponentially growing and $K_n(\xi)$ exponentially dropping at $\xi \rightarrow \infty$:

$$I_n(\xi) \rightarrow \left(\frac{1}{2\pi\xi} \right)^{1/2} e^{\xi}, \quad K_n(\xi) \rightarrow \left(\frac{\pi}{2\xi} \right)^{1/2} e^{-\xi}. \quad (2.158)$$

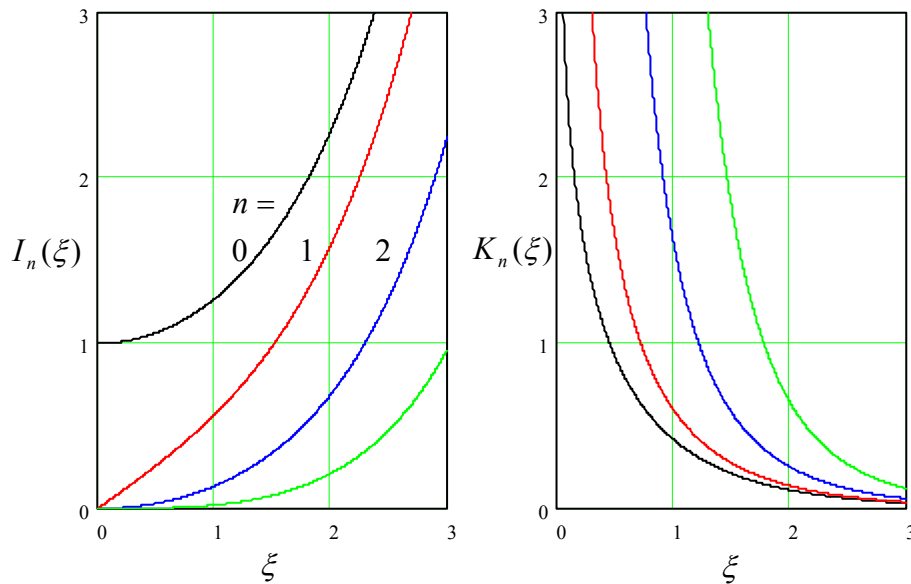


Fig. 2.20. Modified Bessel functions of the first kind (left panel) and the second kind (right panel).

To complete our brief survey of the Bessel functions, let me note that all the functions we have discussed so far may be considered as particular cases of *Bessel functions of the complex argument*, say $J_n(z)$ and $Y_n(z)$, or, alternatively, $H_n^{(1,2)}(z) = J_n(z) \pm iY_n(z)$.⁵¹ The “usual” Bessel functions $J_n(\xi)$ and

⁵¹ These complex functions still obey the general relations (143) and (146), with ξ replaced with z .

$Y_n(\xi)$ may be considered as a set of values of these generalized functions on the real axis ($z = \xi$), while the modified functions as their particular case at $z = i\xi$:

$$I_\nu(\xi) = i^{-\nu} J_\nu(i\xi), \quad K_\nu(\xi) = \frac{\pi}{2} i^{\nu+1} H_\nu^{(1)}(i\xi). \quad (2.159)$$

Moreover, this generalization of the Bessel functions to the whole complex plane z enables the use of their values along other directions on that plane, for example under angles $\pi/4 \pm \pi/2$. As a result, one arrives at the so-called *Kelvin functions*

$$\begin{aligned} \text{ber}_\nu \xi + i \text{bei}_\nu \xi &\equiv J_\nu(\xi e^{-i\pi/4}), \\ \text{ker}_\nu \xi + i \text{kei}_\nu \xi &\equiv i \frac{\pi}{2} H_\nu^{(1)}(\xi e^{-i3\pi/4}), \end{aligned} \quad (2.160)$$

which are also useful for some important problems of mathematical physics and engineering. Unfortunately, we do not have time to discuss these problems in this course.⁵²

(iv) Spherical coordinates are very important in physics, because of the (approximate) spherical symmetry of many objects - from electrons and nuclei and atoms to planets and stars. Let us again require each component ϕ_k of Eq. (84) to satisfy the Laplace equation. Using the well known expression for the Laplace operator in spherical coordinates,⁵³ we get

$$\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial \phi_k}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left(\sin \theta \frac{\partial \phi_k}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2 \phi_k}{\partial \varphi^2} = 0. \quad (2.161)$$

Let us look for a solution of this equation in the following variable-separated form:

$$\phi_k = \frac{\mathcal{R}(r)}{r} \mathcal{P}(\cos \theta) \mathcal{Z}(\varphi), \quad (2.162)$$

Separating equations one by one, just like this has been done in cylindrical coordinates, we get the following equations for the functions participating in this solution:

$$\frac{d^2 \mathcal{R}}{dr^2} - \frac{l(l+1)}{r^2} \mathcal{R} = 0, \quad (2.163)$$

$$\frac{d}{d\xi} \left[(1 - \xi^2) \frac{d\mathcal{P}}{d\xi} \right] + \left[l(l+1) - \frac{\nu^2}{1 - \xi^2} \right] \mathcal{P} = 0, \quad (2.164)$$

$$\frac{d^2 \mathcal{Z}}{d\varphi^2} + \nu^2 \mathcal{Z} = 0, \quad (2.165)$$

where $\xi \equiv \cos \theta$ is a new variable in lieu of θ (so that $-1 \leq \xi \leq +1$), and ν^2 and $l(l+1)$ are the separation constants. (The reason for selection of the latter one in this form will be clear in a minute.) One can see that, in contrast with the cylindrical coordinates, the equation for the radial functions is quite simple.

⁵² Later in the course we will also run into the so-called *spherical Bessel functions* $j_n(\xi)$ and $y_n(\xi)$, which may be expressed via the Bessel functions of a semi-integer order. Surprisingly enough, the spherical Bessel functions turn out to be much simpler than $J_n(\xi)$ and $Y_n(\xi)$.

⁵³ See, e.g., MA Eq. (10.9).

Indeed, let us look for its solution in the form cr^α - just as we have done with Eq. (106). Plugging this solution into Eq. (163), we immediately get the following condition on parameter α :

$$\alpha(\alpha - 1) = l(l + 1). \quad (2.166)$$

This quadratic equation has two roots, $\alpha = l + 1$ and $\alpha = -l$, so that the general solution to Eq. (163) is

$$\mathcal{R} = a_l r^{l+1} + \frac{b_l}{r^l}. \quad (2.167)$$

Equation (165) is also very simple, and to some extent similar to Eq. (108) for the cylindrical coordinates. However, Eq. (164) function $\mathcal{P}(\xi)$, where ξ is the cosine of the polar angle θ , is the so-called *Legendre differential equation*, whose solution cannot be expressed via what is usually called “elementary functions” - though, again, there is no generally accepted line between them and “special functions”.

Let us start with *axially-symmetric problems* for which $\partial\phi/\partial\varphi = 0$. This means $\mathcal{A}(\varphi) = \text{const}$, and thus $\nu = 0$, so that Eq. (164) is reduced to so-called *Legendre's ordinary differential equation*:

$$\frac{d}{d\xi} \left[(1 - \xi^2) \frac{d\mathcal{P}}{d\xi} \right] + l(l + 1)\mathcal{P} = 0. \quad (2.168)$$

Legendre
equation
and
polynomials

One can readily check that the solutions of this equation for integer values of l are specific (*Legendre*) polynomials⁵⁴ that may be defined, for example, by the following *Rodrigues' formula*:

$$\mathcal{P}_l(\xi) = \frac{1}{2^l l!} \frac{d^l}{d\xi^l} (\xi^2 - 1)^l, \quad l = 0, 1, 2, \dots \quad (2.169)$$

As follows from this formula, the first few Legendre polynomials are pretty simple:

$$\begin{aligned} \mathcal{P}_0(\xi) &= 1, \\ \mathcal{P}_1(\xi) &= \xi, \\ \mathcal{P}_2(\xi) &= \frac{1}{2}(3\xi^2 - 1), \\ \mathcal{P}_3(\xi) &= \frac{1}{2}(5\xi^3 - 3\xi), \\ \mathcal{P}_4(\xi) &= \frac{1}{8}(35\xi^4 - 30\xi^2 + 3), \dots \end{aligned} \quad (2.170)$$

though such explicit expressions become more and more bulky as l is increased. As Fig. 21 shows, all these functions, that are defined on the $[-1, +1]$ segment, start at one point, $\mathcal{P}(+1) = +1$, and end up either at the same point or in the opposite point: $\mathcal{P}(-1) = (-1)^l$. On the way between these two end points, the l -th polynomial crosses the horizontal axis l times. It is straightforward to use Eq. (169) for proving that these polynomials form a full, orthogonal set of functions, with the following normalization rule:

⁵⁴ Just for reader's reference: if l is not integer, the general solution of Eq. (2.168) may be represented as a linear combination of the so-called *Legendre functions* (not polynomials!) *of the first and second kind*, $\mathcal{P}(\xi)$ and $\mathcal{Q}(\xi)$.

$$\int_{-1}^{+1} \mathcal{P}_l(\xi) \mathcal{P}_{l'}(\xi) d\xi = \frac{2}{2l+1} \delta_{ll'}, \quad (2.171)$$

so that any function $f(\xi)$, defined on the segment $[-1, +1]$, may be presented as a unique series over the polynomials.⁵⁵

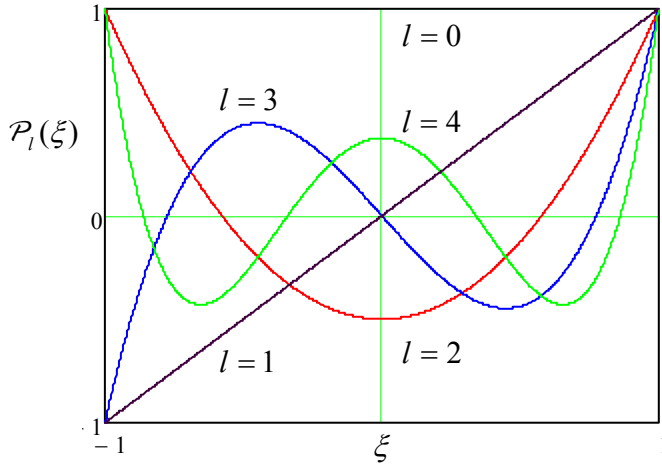


Fig. 2.21. A few lowest Legendre polynomials $\mathcal{P}_l(\xi)$.

Thus, taking into account the additional division by r in Eq. (162), the general solution of any axially-symmetric Laplace problem may be presented as

$$\phi(r, \theta) = \sum_{l=0}^{\infty} \left(a_l r^l + \frac{b_l}{r^{l+1}} \right) \mathcal{P}_l(\cos \theta). \quad (2.172)$$

Variable
separation
in spherical
coordinates
(for axial
symmetry)

Please note a strong similarity between this solution and Eq. (112) for the 2D Laplace problem in polar coordinates. However, besides the difference in angular functions, there is also a difference (by one) in the power of the second radial function, and this difference immediately shows up in core problems.

Indeed, let us solve a problem similar to that shown in Fig. 13: find the electric field around a conducting sphere of radius R , placed into an initially uniform external field \mathbf{E}_0 (whose direction we will take for axis z) – see Fig. 22a. If we select $\phi|_{z=0} = 0$, then $a_0 = b_0 = 0$. Now, just as has been argued for the cylindrical case, at $r \gg R$ the potential should approach that for the uniform field:

$$\phi \rightarrow -E_0 z = -E_0 r \cos \theta, \quad (2.173)$$

and this again means that in Eq. (172), only one of coefficients a_l survives: $a_l = -E_0 \delta_{l1}$. Now, and from the boundary condition on the surface, $\phi(R, \theta) = 0$, we get:

$$0 = \left(-E_0 R + \frac{b_1}{R^2} \right) \cos \theta + \sum_{l \geq 2} \frac{b_l}{R^{l+1}} \mathcal{P}_l(\cos \theta). \quad (2.174)$$

This expression may be viewed as the expansion of function $f(\xi) \equiv 0$ into a series of orthogonal functions $\mathcal{P}_l(\xi)$. Since such expansions are unique, and Eq. (174) is satisfied if

⁵⁵ As a result, there is not practical sense, at least for the purposes of this course, in pursuing (more complex) solutions to Eq. (168) for non-integer values of l .

$$b_l = E_0 R^3 \delta_{l,1}, \quad (2.175)$$

this is indeed the only possibility to satisfy the boundary condition, so that, finally,

$$\phi = -E_0 \left(r - \frac{R^3}{r^2} \right) \cos \theta. \quad (2.176)$$

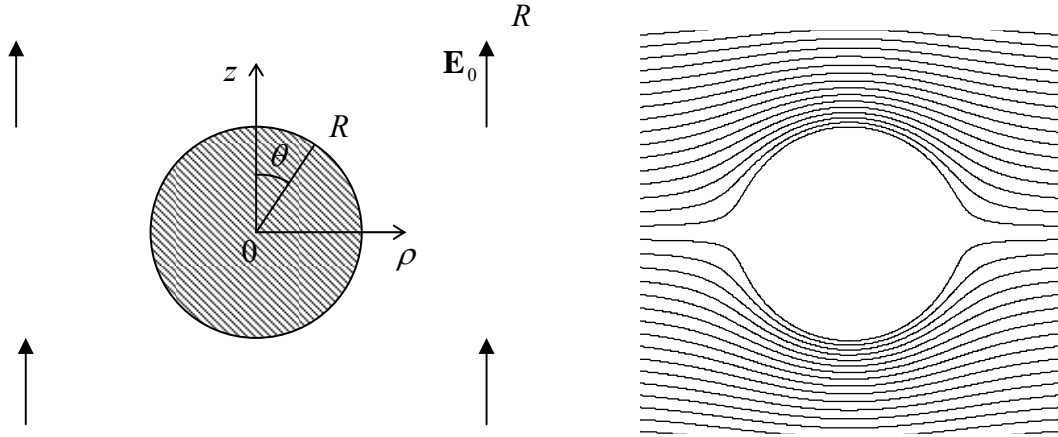


Fig. 2.22. Conducting sphere in a uniform electric field: (a) problem' geometry, and (b) the equipotential surface pattern given by Eq. (176). The pattern is qualitatively similar but quantitatively different from that for the conducting cylinder in a perpendicular field – cf. Fig. 13.

This distribution, shown in Fig. 22b, is very much similar to Eq. (117) for the cylindrical case, but with a different power of radius in the second term. This leads to a quantitatively different distribution of the surface electric field:

$$E_n = -\frac{\partial \phi}{\partial r} \Big|_{r=R} = 3E_0 \cos \theta, \quad (2.177)$$

so that its maximal value is a factor of 3 (rather than 2) larger than the external field.

Now let us discuss the Laplace equation solution in the general case (no axial symmetry), but only for most important systems in which the free space surrounds the origin from all sides. In this case the solutions to Eq. (165) have to be 2π -periodic, and hence $\nu = n = 0, \pm 1, \pm 2, \dots$. Mathematics says that the Legendre equation (164) with integer $\nu = n$ and a fixed integer l has a solution only for a limited range of n :⁵⁶

$$-l \leq n \leq +l. \quad (2.178)$$

These solutions are called the *associated Legendre functions*. For $n \geq 0$, they may be defined via the Legendre polynomials using the following formula:

⁵⁶ In quantum mechanics, letter n is typically reserved used for the “main quantum number”, while the azimuthal functions are numbered by index m . However, I will keep using n as their index, because for this course's purposes, this seems more logical in the view of the similarity of the spherical and cylindrical functions.

$$\mathcal{P}_l^n(\xi) = (-1)^n (1 - \xi^2)^{n/2} \frac{d^n}{d\xi^n} \mathcal{P}_l(\xi). \quad (2.179)$$

On the segment $\xi \in [-1, +1]$, each set of the associated Legendre functions with a fixed index n and non-negative l form a full, orthogonal set, with the normalization relation,

$$\int_{-1}^{+1} \mathcal{P}_l^n(\xi) \mathcal{P}_{l'}^n(\xi) d\xi = \frac{2}{2l+1} \frac{(l+n)!}{(l-n)!} \delta_{ll'}, \quad (2.180)$$

that is evidently a generalization of Eq. (171).

Since these relations may seem a bit intimidating, let me write down explicit expressions for a few $\mathcal{P}_n^l(\cos\theta)$ with the lowest values of l and $n \geq 0$:

$$l = 0: \quad \mathcal{P}_0^0(\cos\theta) = 1; \quad (2.181)$$

$$l = 1: \quad \begin{cases} \mathcal{P}_1^0(\cos\theta) = \cos\theta, \\ \mathcal{P}_1^1(\cos\theta) = -\sin\theta; \end{cases} \quad (2.182)$$

$$l = 2: \quad \begin{cases} \mathcal{P}_2^0(\cos\theta) = (3\cos^2\theta - 1)/2, \\ \mathcal{P}_2^1(\cos\theta) = -2\sin\theta\cos\theta, \\ \mathcal{P}_2^2(\cos\theta) = -3\cos^2\theta. \end{cases} \quad (2.183)$$

The reader should agree there is not much intimidation in these functions - which are most important for applications.

Now the general solution (162) to the Laplace equation in the spherical coordinates may be spelled out as

$$\phi(r, \theta, \varphi) = \sum_{l=0}^{\infty} \left(a_l r^l + \frac{b_l}{r^{l+1}} \right) \sum_{n=0}^l \mathcal{P}_l^n(\cos\theta) \mathcal{Z}_n(\varphi), \quad \mathcal{Z}_n(\varphi) = c_n \cos n\varphi + s_n \sin n\varphi. \quad (2.184)$$

Variable
separation
in spherical
coordinates
(general
case)

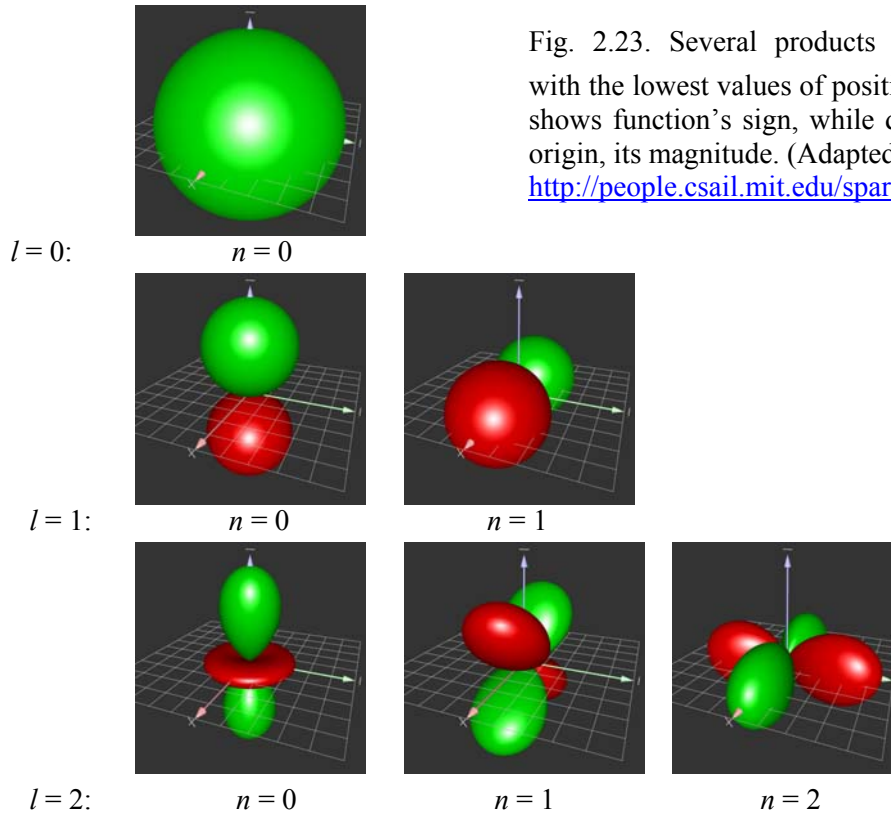
Since the difference between angles θ and φ is somewhat artificial, physicists prefer to think not about functions \mathcal{P} and \mathcal{Z} in separation, but directly about their products that participate in this solution. Figure 23 shows a few such angular functions⁵⁷ by plotting their modulus along the radius, and using bi-color to show the function sign. While the lowest function ($l = 0, n = 0$) is just a constant, two “dipole” functions ($l = 1$) differ from each other by their spatial orientation. Functions with higher l (say, $l = 2$) differ more substantially, with the following general trend: for each value of l , the function with $n = 0$ is

⁵⁷ In quantum mechanics, it is more convenient to use a slightly different set of basic functions, namely complex functions called *spherical harmonics*,

$$Y_l^n(\theta, \varphi) \equiv \left[\frac{2l+1}{4\pi} \frac{(l-n)!}{(l+n)!} \right]^{1/2} \mathcal{P}_l^n(\cos\theta) e^{in\varphi},$$

which are defined for both positive and negative n (within the limits $-l \leq n \leq +l$), because they form a full set of orthonormal eigenfunctions of angular momentum operators L^2 and L_z - see, e.g., QM Secs. 3.6 and 5.6.

axially-symmetric⁵⁸ and has l zeros on its way from $\theta = 0$ to $\theta = \pi$, while the functions with $n = l$ do not have zeros inside that interval, while oscillating most strongly as functions of φ .



As an exception, in order to save time, I will skip an example of application of the associated Legendre functions, because several such examples are given in the quantum mechanics part of these series. (Note that in this field, index n is traditionally called m – the *magnetic quantum number*.)

2.6. Charge images

So far, we have discussed various methods of solution of the *Laplace* boundary problem (35). Let us now move on to the discussion of its generalization, the *Poisson* equation (1.41), that we need when besides the conductors, we also have “free” charges with a known spatial distribution $\rho(\mathbf{r})$. (This will also allow us, better equipped, to revisit the Laplace problem again in the next section.)

Let us start with a somewhat limited, but sometimes very useful *charge image* (or “image charge”) *method*. Consider a very simple problem: a single point charge near a conducting half-space – see Fig. 24. Let us prove that its solution, above conductor's surface ($z \geq 0$), may be presented as:

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \left(\frac{q}{r_1} - \frac{q}{r_2} \right) = \frac{q}{4\pi\epsilon_0} \left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} - \frac{1}{|\mathbf{r} - \mathbf{r}''|} \right), \quad (2.185)$$

⁵⁸ According to Eq. (179), these functions involve only the Legendre polynomials $\mathcal{P}_l \equiv \mathcal{P}_l^0$.

or in a more explicit (coordinate) form:

$$\phi(\mathbf{r}) = \frac{q}{4\pi\epsilon_0} \left(\frac{1}{[\rho^2 + (z-d)^2]^{1/2}} - \frac{1}{[\rho^2 + (z+d)^2]^{1/2}} \right), \quad (2.186)$$

where ρ is the distance of the observation point from the vertical line on which the charge is located. Indeed, this solution evidently satisfies both the boundary condition of zero potential at the surface of the conductor ($z = 0$), and the Poisson equation (1.41), with the single δ -functional source at point $\mathbf{r}' = \{0, 0, d\}$ in its right-hand part, because its another singularity, at point $\mathbf{r}'' = \{0, 0, -d\}$, is outside the region of validity of this solution ($z \geq 0$).

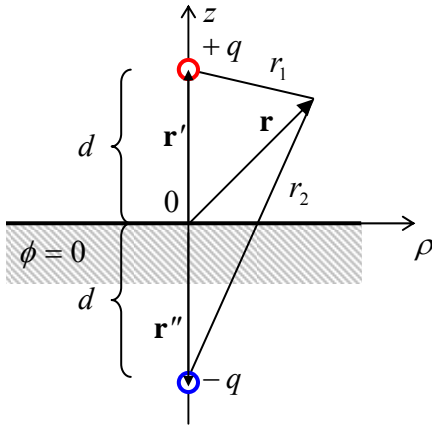


Fig. 2.24. The simplest problem readily solvable by the charge image method. Point colors in this section are used, here and in the balance of this section, to denote charges of the original (red) and opposite (blue) sign.

Physically, the solution may be interpreted as the sum of the fields of the actual charge ($+q$) at point \mathbf{r}' , and an equal but opposite charge ($-q$) at the “mirror image” point \mathbf{r}'' (Fig. 24). This is the basic idea of the charge image method. Before moving to more complex problems, let us discuss the situation shown in Fig. 24 in a little bit more detail. First, we can use Eqs. (3) and (186) to calculate the surface charge density:

$$\sigma = -\epsilon_0 \left. \frac{\partial \phi}{\partial z} \right|_{z=0} = -\frac{q}{4\pi} \frac{\partial}{\partial z} \left(\frac{1}{[\rho^2 + (z-d)^2]^{1/2}} - \frac{1}{[\rho^2 + (z+d)^2]^{1/2}} \right)_{z=0} = -\frac{q}{4\pi} \frac{2d}{(\rho^2 + d^2)^{3/2}}. \quad (2.187)$$

The total surface charge is

$$Q = \int_A \sigma d^2r = 2\pi \int_0^\infty \sigma(\rho) \rho d\rho = -\frac{q}{2} \int_0^\infty \frac{d}{(\rho^2 + d^2)^{3/2}} 2\rho d\rho. \quad (2.188)$$

This integral may be easily taken using the substitution $\xi \equiv \rho^2/d^2$ (giving $d\xi = 2\rho d\rho/d^2$):

$$Q = -\frac{q}{2} \int_0^\infty \frac{d\xi}{(\xi + 1)^{3/2}} = -q. \quad (2.189)$$

This result is very natural, because the conductor “wants” to bring as much surface charge from its interior to the surface as necessary to fully compensate the initial charge ($+q$) and hence to kill the

electric field at large distances as efficiently as possible, hence reducing the total electrostatic energy (1.67) to the lowest possible value.

For a deeper understanding of this *polarization charge* of the surface, let us take our calculations to the extreme – to q equal to one elementary charge e , and place a particle with this charge (for example, a proton) at a macroscopic distance – say 1 m – from conductor's surface. Then, according to Eq. (189), the total polarization charge of the surface equals to that of an electron, and according to Eq. (187), its spatial extent is of the order of $d^2 = 1 \text{ m}^2$. This means that if we consider a much smaller part of the surface, $\Delta A \ll d^2$, its polarization charge magnitude $\Delta Q = \sigma \Delta A$ is much *less than one electron*! For example, Eq. (187) shows that the polarization charge of quite a macroscopic area $\Delta A = 1 \text{ cm}^2$ right under the initial charge ($\rho = 0$) is $e\Delta A/2\pi d^2 \approx 1.6 \times 10^{-5} e$. Can this be true, or our theory is somehow limited to the charges much larger than e ?

Surprisingly enough, the answer to this question has become clear (at least to some physicists :-)) only as late as in the mid-1980s when several experiments demonstrated, and theorists accepted, some rather grudgingly that the usual polarization charge formulas are valid for elementary charges q as well, i.e., such the polarization charge ΔQ of a macroscopic surface area can indeed be less than e . The underlying reason for this paradox is the nature of the polarization charge of the conductor surface: as should be clear from our discussion in Sec. 1, it is due not to new charged particles brought into the conductor (such charge would be in fact quantized in the units of e), but to a small *shift* of the free charges of a conductor by a very small distance from their equilibrium positions that they had in the absence of the external field induced by charge q . This shift is not quantized, at least on the scale relevant for our issue, and neither is ΔQ . This understanding has opened a way toward the invention and experimental demonstration of several new devices including so-called *single-electron transistors*,⁵⁹ which may be, in particular, used to measure polarization charges as small as $\sim 10^{-6} e$.

To complete the discussion of our initial problem (Fig. 24), let us find the potential energy U of the charge-to-surface interaction. For that we may use the value of the electrostatic potential (185) in the point of the charge itself ($\mathbf{r} = \mathbf{r}'$), of course ignoring the infinite potential created by the charge itself, so that the remaining potential is that of the image charge

$$\phi_{\text{image}}(\mathbf{r}') = -\frac{1}{4\pi\epsilon_0} \frac{q}{2d}. \quad (2.190)$$

Looking at the definition of the electrostatic potential, given by Eq. (1.31), it may be tempting to immediately write $U = q\phi_{\text{image}} = - (1/4\pi\epsilon_0)(q^2/2d)$ **[WRONG!]**, but this would not be correct. The reason is that potential ϕ_{image} is not independent of q , but is actually induced by this charge. This is why the correct approach is to use Eq. (1.63), with just one term:

$$U = \frac{1}{2} q \phi_{\text{image}} = -\frac{1}{4\pi\epsilon_0} \frac{q^2}{4d}, \quad (2.191)$$

twice lower in magnitude than the wrong result cited above. In order to double-check this result, and also get a better feeling of the factor $1/2$ that distinguishes it from the wrong guess, we can recalculate

⁵⁹ Actually, this term (for which the author of these notes should be blamed :-)) is misleading: operation of the “single-electron transistor” is based on the interplay of discrete charges (multiples of e) transferred between conductors, and *sub*-single-electron polarization charges – see, e.g., K. K. Likharev, *Proc. IEEE* **87**, 606 (1999).

energy U as the integral of the force exerted on the charge by the conductor (i.e., in our formalism, by the image charge):

$$U = -\int_{\infty}^d F(z) dz = \frac{1}{4\pi\epsilon_0} \int_{\infty}^d \frac{q^2}{(2z)^2} dz = -\frac{1}{4\pi\epsilon_0} \frac{q^2}{4d}. \quad (2.192)$$

This calculation clearly accounts for the gradual build-up of force F , as the real charge is brought from afar (where we have opted for $U=0$) toward the surface.

This result, used for electrons, particles with charge $q = -e$, has several important applications. For example, let us plot energy U for an electron near a metallic surface, as a function of d . For that, we may use Eq. (192) until our macroscopic approximation (2) becomes invalid, and U transitions to some negative constant value ($-\psi$) inside the conductor – see Fig. 25a.

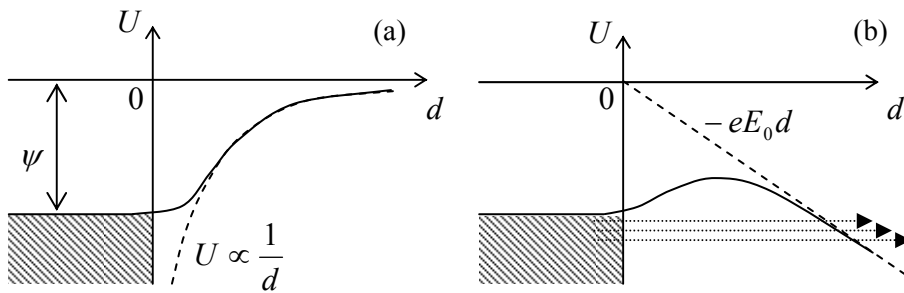


Fig. 2.25. (a) Origin of the workfunction and (b) the field emission of electrons (schematically).

The positive constant ψ is called *workfunction*, because it describes how much work should be done on an electron to remove it from the conductor. As was discussed in Sec. 1, in good metals the electric field screening happens at interatomic distances $a_0 \approx 10^{-10}$ m. Plugging $d = a$ and $q = -e$ into Eq. (191), we get $\approx 6 \times 10^{-19}$ J ≈ 3.5 eV. This crude estimate is in a surprisingly good agreement with the experimental values of the workfunction, ranging between 4 and 5 eV for most metals.⁶⁰

Next, let us consider the effect of an additional external electric field \mathbf{E}_0 applied perpendicular to a metallic surface, on this potential profile. Assuming the field to be uniform, we can add its potential $-eE_0d$ at distance d from the surface, to that created by the image charge. (As we know from Eq. (1.53), since field \mathbf{E}_0 is independent of the electron position, its recalculation to the potential energy does not require the coefficient $1/2$.) As the result, the potential energy of an electron near the surface becomes

$$U(d) = -eE_0d - \frac{1}{4\pi\epsilon_0} \frac{e^2}{4d}, \quad \text{for } d > a_0, \quad (2.193)$$

with a similar crossover to $U = -\psi$ inside the conductor – see Fig. 25b. One can see that at the appropriate sign, and sufficient magnitude of the applied field, it lowers the potential barrier that prevented electron from leaving the conductor. At $E_0 \sim \psi/a_0$ this suppression becomes so strong that electrons just below the Fermi surface start quantum-mechanical tunneling through the remaining thin

⁶⁰ For more discussion of workfunction, and its effect on electron kinetics, see, e.g., SM Sec. 6.4.

barrier. This is the *field emission* effect, which is used in vacuum electronics to provide efficient cathodes that do not require heating to high temperatures.⁶¹

Returning to the basic electrostatics, let us consider some other geometries where the method of images may be effectively applied. First, let us consider a right corner (Fig. 26a). Reflecting the initial charge in the vertical plane we get the image charge shown in the top left corner of the panel, that makes the boundary condition $\phi = \text{const}$ satisfied on the vertical surface of the corner. However, in order the same to be true on the horizontal surface, we have to reflect *both* the initial charge *and* the image charge in the horizontal plane, flipping their signs. The final configuration of 4 charges, shown in Fig. 26a, satisfies all the boundary conditions. The resulting potential distribution may be readily written as the evident generalization of Eq. (185). From there, the electric field and electric charge distributions, and the potential energy and forces acting on the charge may be calculated exactly as above.

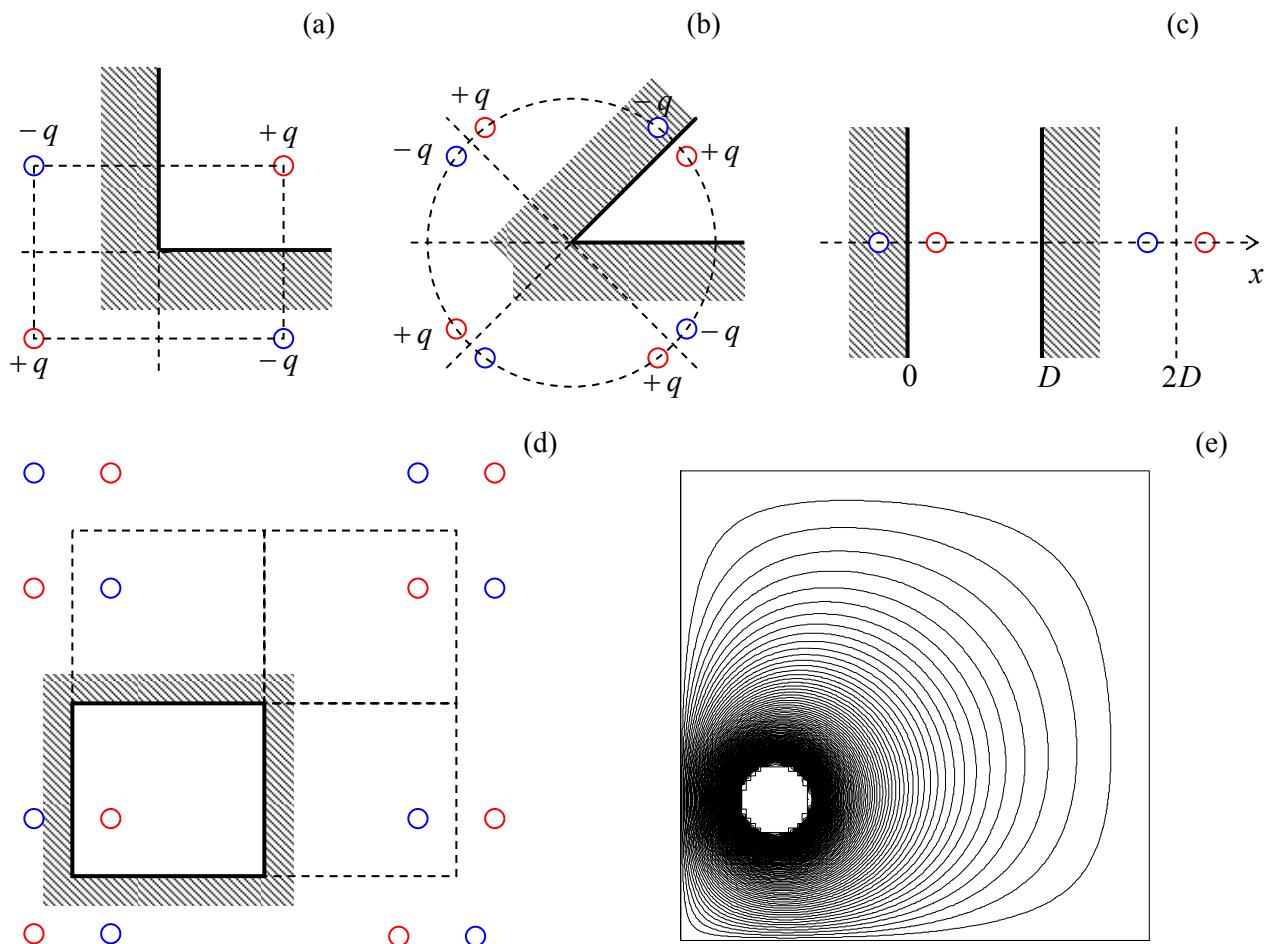


Fig. 2.26. Charge images for (a, b) internal corners with angles π and $\pi/2$, (c) plane capacitor, and (d) rectangular box, and (e) equipotential surfaces for the last system.

⁶¹ The practical development of such “cold” cathodes is strongly affected by the fact that any nanoscale surface irregularity (a protrusion, an atomic cluster, or even a single “adatom” stuck to the surface) may cause a strong increase of the local field well above the average applied field E_0 (see, for example our discussion in Sec. 4 above), making the emission reproducibility an issue.

Next, consider a corner with angle $\pi/4$ (Fig. 26b). Here we need to repeat the reflection operation not 2 but 4 times before we arrive at the final pattern of 8 positive and negative charges. (Any attempt to continue this process would lead to an overlap with the already existing charges.) This reasoning can be readily extended to any 2D corner with angle $\beta = \pi/n$, with any integer n , that requires $2n$ charges (including the initial one) to satisfy all the boundary conditions.

Some configurations require an infinite number of images that are, however, tractable. The most important of them is a system of two parallel conducting surfaces, i.e. a plane capacitor of infinite area (Fig. 26c). Here the repeated reflection leads to an infinite system of charges $\pm q$ at points

$$x_j^\pm = \pm d + 2Dj, \quad (2.194)$$

where $0 < d < D$ is the position of the initial charge and j an arbitrary integer. However, the resulting infinite series for the potential of the real charge q , created by the field of its images,

$$\phi(d) = \frac{1}{4\pi\epsilon_0} \left[-\frac{q}{2d} + \sum_{j \neq 0} \sum_{\pm} \frac{\pm q}{|d - x_j^\pm|} \right] = -\frac{q}{4\pi\epsilon_0} \left[\frac{1}{2d} + \frac{d^2}{D^3} \sum_{j=1}^{\infty} \frac{1}{j[j^2 - (d/D)^2]} \right], \quad (2.195)$$

is converging (in its last form) very fast. For example, the exact value, $\phi(D/2) = -2\ln 2(q/4\pi\epsilon_0 D)$, differs by less than 5% from the approximation using just the first term of the sum.

The same method may be applied to 2D (cylindrical) and 3D rectangular boxes that require, respectively, a 2D or 3D infinite lattices of images; for example in a 3D box with sides a , b , and c , charges $\pm q$ are located at points (Fig. 26d)

$$\mathbf{r}_{jkl}^\pm = \pm \mathbf{r}' + 2ja + 2kb + 2lc, \quad (2.196)$$

where \mathbf{r}' is the location of the initial (real) charge, and j , k , and l are arbitrary integers. Figure 26e shows the results of summation of the potentials of such charge set, including the real one, in a 2D box (within the plane of the real charge). One can see that the equipotential surfaces, concentric near the charge, are naturally leaning along the conducting walls of the box, which should be equipotential.

Even more surprisingly, the image charge method works very efficiently not only for the rectilinear geometries, but also for spherical ones. Indeed, let us consider a point charge q at some distance d from the center of a conducting, grounded sphere of radius R (Fig. 27a), and try to satisfy the boundary condition $\phi = 0$ for the electrostatic potential on sphere's surface using an imaginary charge q' located at some point located beyond the surface, i.e. inside the sphere.

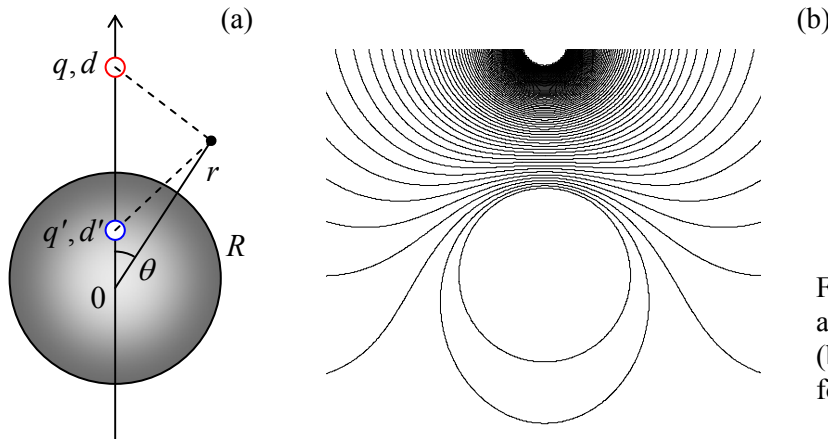


Fig. 2.27. Method of charge images for a conducting sphere: (a) the idea, and (b) the resulting potential distribution for particular case $d = 2R$.

From problem's symmetry, it is clear that the point should be at the line passing through the real charge and the sphere's center, at some distance d' from the center. Then the total potential created by the two charges at an arbitrary point with $r \geq R$ (Fig. 27a) is

$$\phi(r, \theta) = \frac{1}{4\pi\epsilon_0} \left[\frac{q}{(r^2 + d^2 - 2rd \cos \theta)^{1/2}} + \frac{q'}{(r^2 + d'^2 - 2rd' \cos \theta)^{1/2}} \right]. \quad (2.197)$$

It is easy to see that we can make the two fractions to be equal and opposite at all points on the sphere's surface (i.e. for any θ at $r = R$), if we take⁶²

$$d' = \frac{R^2}{d}, \quad q' = -\frac{R}{d} q. \quad (2.198)$$

Since the solution to any Poisson boundary problem is unique, Eqs. (197) and (198) give us such solution for this problem. Figure 27b shows a typical equipotential pattern calculated using Eqs. (197) and (198). It is surprising how formulas that simple may describe such a nontrivial field distribution.

Now let us calculate the total charge Q induced by charge q on conducting sphere's surface. We could do this, as we have done for the conducting plane, by the brute force integration of the surface charge density $\sigma = -\epsilon_0 \partial \phi / \partial r|_{r=R}$. It is more elegant, however, to use the following Gauss law argument. Expression (197) is valid (at $r \geq R$) regardless whether we are dealing with our real problem (charge q and the conducting sphere) or with the equivalent charge configuration (point charges q and q' , with no sphere at all). Hence, according to Eq. (1.16), the Gaussian integral over a surface with radius $r = R + 0$, and the total charge inside the sphere should be also the same. Hence we immediately get

$$Q = q' = -\frac{R}{d} q. \quad (2.199)$$

The similar argumentation may be used to find the charge-to-sphere interaction force:

$$F = qE_{\text{image}}(d) = q \frac{q'}{4\pi\epsilon_0 (d - d')^2} = -\frac{q^2}{4\pi\epsilon_0} \frac{R}{d} \frac{1}{(d - R^2/d)^2} = -\frac{q^2}{4\pi\epsilon_0} \frac{Rd}{(d^2 - R^2)^2}. \quad (2.200)$$

(Note that this expression is legitimate only at $d > R$.) At large distances, $d/R \gg 1$, this attractive force decreases as $1/d^3$. This unusual dependence arises because, as Eq. (198) specifies, the induced charge of the sphere, responsible for the force, is not constant but decreases as $Q \propto 1/d$.

All the previous formulas referred to a sphere that is grounded to keep its potential equal to zero. But what if we keep the sphere galvanically insulated, so that its net charge is fixed, e.g., equals zero? Instead of solving the problem from the scratch, let us use (again!) the linear superposition principle. For that, we may add to the previous problem an additional charge, equal to $(-Q)$, to the sphere, and argue that this addition gives an additional potential that does not depend of the potential induced by charge q . For the interaction force, such addition yields

⁶² In geometry, such points, with $dd' = R^2$, are referred to as the result of mutual *inversion* in a sphere of radius R .

$$F = \frac{qq'}{4\pi\epsilon_0(d-d')^2} + \frac{qQ}{4\pi\epsilon_0 d^2} = -\frac{q^2}{4\pi\epsilon_0} \left[\frac{Rd}{(d^2 - R^2)^2} - \frac{R}{d^3} \right]. \quad (2.201)$$

At large distances, the two terms proportional to $1/d^3$ cancel each other, giving $F \propto 1/d^5$. Such a rapid force decay is due to the fact that the field of the uncharged sphere is equivalent to that of two (equal and opposite) induced charges $+Q$ and $-Q$, and the distance between them ($d' = R^2/d$) tends to zero at $d \rightarrow \infty$. The potential energy of such interaction behaves as $U \propto 1/d^6$ at $d \rightarrow \infty$; in the next chapter we will see that this is the general law of the induced dipole interaction.

2.7. Green's functions

I have spent so much time/space discussing the potential distributions created by a single point charge in various conductor geometries, because, for any geometry, the generalization of these results to the arbitrary distribution $\rho(\mathbf{r})$ of free charges is straightforward. Namely, if a single charge q , located at point \mathbf{r}' , created electrostatic potential

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} q G(\mathbf{r}, \mathbf{r}'), \quad (2.202)$$

then, due to the linear superposition principle, an arbitrary charge distribution creates potential

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \sum_j q_j G(\mathbf{r}, \mathbf{r}_j) = \frac{1}{4\pi\epsilon_0} \int \rho(\mathbf{r}') G(\mathbf{r}, \mathbf{r}') d^3 r'. \quad (2.203)$$

Spatial
Green's
function

Kernel $G(\mathbf{r}, \mathbf{r}')$ is called the (spatial) *Green's function* – the notion very popular in all fields of physics.⁶³ Evidently, as Eq. (1.35) shows, in the unlimited free space

$$G(\mathbf{r}, \mathbf{r}') = \frac{1}{|\mathbf{r} - \mathbf{r}'|}, \quad (2.204)$$

i.e. the Green's function depends only on one scalar argument – the distance between the field observation point \mathbf{r} and the field-source (charge) point \mathbf{r}' . However, as soon as there are conductors around, the situation changes. In this course we will only deal with Green's functions that are defined in the space between conductors, and that vanish as soon as the radius-vector \mathbf{r} points to the surface of any conductor:⁶⁴

$$G(\mathbf{r}, \mathbf{r}')|_{\mathbf{r} \in A} = 0. \quad (2.205)$$

With this definition, it is straightforward to deduce the Green's functions for the solutions of the last section's problems in which conductors were grounded ($\phi = 0$). For example, for a semi-space $z \geq 0$ limited by a conducting plane (Fig. 24), Eq. (185) yields

$$G = \frac{1}{|\mathbf{r} - \mathbf{r}'|} - \frac{1}{|\mathbf{r} - \mathbf{r}''|}, \quad \text{with } \rho'' = \rho' \text{ and } z'' = -z'. \quad (2.206)$$

⁶³ See, e.g., CM Sec. 4.1, QM Secs. 2.2, 7.2 and 7.4, and SM Sec. 5.5.

⁶⁴ G so defined is sometimes called the *Dirichlet function*.

We see that in the presence of conductors (and, as we will see later, any other polarizable media), the Green's function may depend not only on the difference $\mathbf{r} - \mathbf{r}'$, but in a specific way from each of these two arguments.

So far, this looked just like re-naming our old results. The really non-trivial result of the Green's function application to electrostatics is that, somewhat counter-intuitively, the knowledge of the Green's function for a system with *grounded* conductors (Fig. 28a) allows one to calculate the field created by *voltage-biased* conductors (Fig. 28b), with the same geometry.

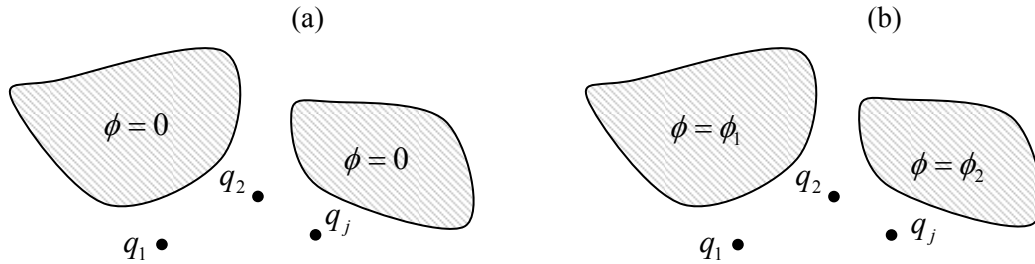


Fig. 2.28. Green's function method allows the solution of a simpler boundary problem (a) to be used to find the solution of a more complex problem (b), for the same conductor geometry.

In order to show that, let us use the so-called *Green's theorem* of the vector calculus.⁶⁵ The theorem states that for any two scalar, differentiable functions $f(\mathbf{r})$ and $g(\mathbf{r})$, and any volume V ,

$$\int_V (f \nabla^2 g - g \nabla^2 f) d^3 r = \oint_S (f \nabla g - g \nabla f)_n d^2 r, \quad (2.207)$$

where S is the surface limiting the volume. Applying the theorem to the electrostatic potential $\phi(\mathbf{r})$ and the Green's function G (also considered as a function of \mathbf{r}), let us use the Poisson equation (1.41) to replace $\nabla^2 \phi$ with $(-\rho/\epsilon_0)$, and notice that G , considered as a function of \mathbf{r} , obeys the Poisson equation with the δ -functional source:

$$\nabla^2 G(\mathbf{r}, \mathbf{r}') = -4\pi \delta(\mathbf{r} - \mathbf{r}'). \quad (2.208)$$

(Indeed, according to its definition (202), this function may be formally considered as the field of a point charge $q = 4\pi\epsilon_0$.) Now swapping the notation of radius-vectors, $\mathbf{r} \leftrightarrow \mathbf{r}'$, and using the Green's function symmetry, $G(\mathbf{r}, \mathbf{r}') = G(\mathbf{r}', \mathbf{r})$,⁶⁶ we get

$$-4\pi\phi(\mathbf{r}) - \int_V \left(-\frac{\rho(\mathbf{r}')}{\epsilon_0} \right) G(\mathbf{r}, \mathbf{r}') d^3 r' = \oint_S \left[\phi(\mathbf{r}') \frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial n'} - G(\mathbf{r}, \mathbf{r}') \frac{\partial \phi(\mathbf{r}')}{\partial n'} \right] d^2 r'. \quad (2.209)$$

Let us apply this relation to volume V of *free space* between the conductors, and the boundary A slightly outside of their surface. In this case, by its definition, the Green's function $G(\mathbf{r}, \mathbf{r}')$ vanishes at the conductor surface ($\mathbf{r} \in S$) – see Eq. (205). Now changing the sign of $\partial n'$ (so that it would be the outer normal for *conductors*, rather than free space volume V), dividing all terms by 4π , and partitioning

⁶⁵ See, e.g., MA Eq. (12.3). Actually, this theorem is a ready corollary of the divergence theorem, MA Eq. (12.2).

⁶⁶ This symmetry, virtually evident from Eq. (204), may be formally proved by applying Eq. (207) to functions $f(\mathbf{r}) \equiv G(\mathbf{r}, \mathbf{r}')$ and $g(\mathbf{r}) \equiv G(\mathbf{r}, \mathbf{r}'')$. With this substitution, the left-hand part becomes equal to $-4\pi[G(\mathbf{r}'', \mathbf{r}') - G(\mathbf{r}', \mathbf{r}'')]$, while the right-hand part is zero, due to Eq. (205).

the total surface A into the parts (numbered by index j) corresponding to different conductors (possibly, kept at different potentials ϕ_k), we finally arrive at the famous result:⁶⁷

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int_V \rho(\mathbf{r}') G(\mathbf{r}, \mathbf{r}') d^3 r' + \frac{1}{4\pi} \sum_k \phi_k \oint_{S_k} \frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial n'} d^2 r'. \quad (2.210)$$

Potential
via Green's
function

While the first term in the right-hand part of this relation is a direct and evident expression of the superposition principle, given by Eq. (203), the second term is highly non-trivial: it describes the effect of conductors with *nonvanishing* potentials ϕ_k (Fig. 28b) using the Green's function calculated for the similar system with *grounded* conductors, i.e. with all $\phi_k = 0$ (Fig. 28a). Let me emphasize that since our volume V excludes conductors, the first term in the right-hand part of Eq. (210) includes only the “free-standing” charges of the system (in Fig. 28, marked q_1, q_2 , etc.), but not the surface charges of the conductors – which are taken into account, implicitly, by the second term.

In order to illustrate what a powerful tool Eq. (210) is, let us use to calculate the electrostatic field in two systems. In the first of them, a circular disk, separated with a very thin cut from a conducting plane, is biased with potential $\phi = V$, while the rest of the plane is grounded - see Fig. 29.

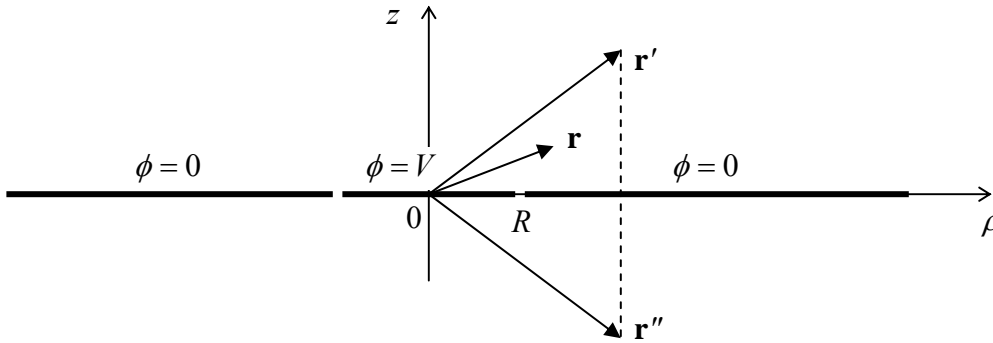


Fig. 2.29. Voltage-biased conducting circle inside a grounded conducting plane.

If the width of the gap between the circle and rest of the plane is negligible, we may apply Eq. (210) with $\rho(\mathbf{r}') = 0$, and the Green's function for the uncut plane – see Eq. (206).⁶⁸ In the cylindrical coordinates, the function may be rewritten as

$$G(\mathbf{r}, \mathbf{r}') = \frac{1}{(\rho^2 + \rho'^2 - 2\rho\rho' \cos(\varphi - \varphi') + (z - z')^2)^{1/2}} - \frac{1}{(\rho^2 + \rho'^2 - 2\rho\rho' \cos(\varphi - \varphi') + (z + z')^2)^{1/2}}. \quad (2.211)$$

(The sum of the first three terms under the square roots of Eq. (211) is just the squared distance between the horizontal projections ρ and ρ' of vectors \mathbf{r} and \mathbf{r}' (or \mathbf{r}''), correspondingly, while the last terms are the squares of their vertical spacings.)

Now we can readily calculate the necessary derivative:

⁶⁷ In some textbooks, the sign before the surface integral is negative, because their authors use the outer normal of the *free-space* region V rather than *that occupied by conductors* - as I do.

⁶⁸ Indeed, if all parts of the cut plane are grounded, a narrow cut does not change the field distribution, and hence the Green's function, significantly.

$$\left. \frac{\partial G}{\partial n'} \right|_s = \left. \frac{\partial G}{\partial z'} \right|_{z'=+0} = \frac{2z}{(\rho^2 + \rho'^2 - 2\rho\rho' \cos(\varphi - \varphi') + z^2)^{3/2}}. \quad (2.212)$$

Due the axial symmetry of the system, we can take φ for zero. With this, Eqs. (210) and (212) yield

$$\phi = \frac{V}{4\pi} \oint_s \frac{\partial G(\mathbf{r}, \mathbf{r}')}{\partial n'} d^2 r' = \frac{Vz}{2\pi} \int_0^{2\pi} d\varphi' \int_0^R \frac{\rho' d\rho'}{(\rho^2 + \rho'^2 - 2\rho\rho' \cos \varphi' + z^2)^{3/2}}. \quad (2.213)$$

This integral is not too pleasing, but may be readily worked out for points on the symmetry axis ($\rho = 0$):

$$\phi = Vz \int_0^R \frac{\rho' d\rho'}{(\rho'^2 + z^2)^{3/2}} = \frac{V}{2} \int_0^{R^2/z^2} \frac{d\xi}{(\xi + 1)^{3/2}} = V \left[1 - \frac{z}{(R^2 + z^2)^{1/2}} \right]. \quad (2.214)$$

This expression shows that if $z \rightarrow 0$, the potential tends to V (as it should), while at $z \gg R$,

$$\phi \rightarrow V \frac{R^2}{2z^2}. \quad (2.215)$$

This asymptotic behavior is typical for electric dipoles – see the next chapter.

Now, let us use the same Eq. (210) to solve the (in :-)-famous problem of the cut sphere (Fig. 30). Again, if the gap between the two conducting semi-spheres is very thin ($t \ll R$), we may use the Green's function for the grounded (and uncut) sphere. For a particular case $\mathbf{r}' = d\mathbf{n}_z$, this function is given by Eqs. (197)-(198); generalizing the former relation for an arbitrary direction of vector \mathbf{r}' , we get

$$G = \frac{1}{(r^2 + r'^2 - 2rr' \cos \gamma)^{1/2}} - \frac{R/r'}{(r^2 + (R^2/r')^2 - 2r(R^2/r') \cos \gamma)^{1/2}}, \quad \text{for } r, r' \geq R, \quad (2.216)$$

where γ is the angle between vectors \mathbf{r} and \mathbf{r}' , and hence \mathbf{r}'' (Fig. 30).

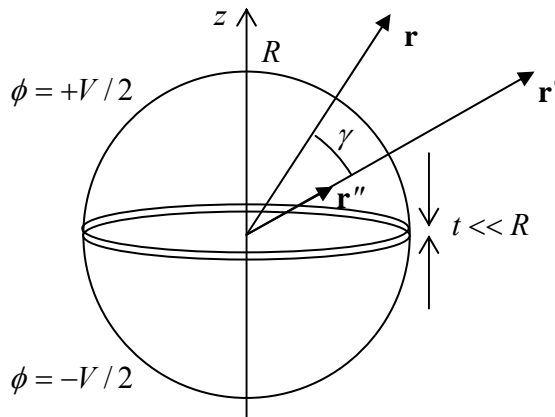


Fig. 2.30. A system of two, oppositely biased semi-spheres.

Now, finding the Green's function's derivative,

$$\left. \frac{\partial G}{\partial r'} \right|_{r'=R+0} = - \frac{(r^2 - R^2)}{R[r^2 + R^2 - 2Rr \cos \gamma]^{3/2}}, \quad (2.217)$$

and plugging it into Eq. (210), we see that the integration is easy only for the field on the symmetry axis ($\mathbf{r} = r\mathbf{n}_z$, $\gamma = \theta$), giving

$$\phi = \frac{V}{2} \left[1 - \frac{z^2 - R^2}{z(z^2 + R^2)^{1/2}} \right]. \quad (2.218)$$

For $z \rightarrow R$, $\phi \rightarrow V/2$ (just checking :-), while for $z \gg R$,

$$\phi \rightarrow V \frac{3R^2}{4z^2}, \quad (2.219)$$

so this is also an electric dipole field – see the next chapter.

2.8. Numerical methods

Despite the richness of analytical methods, for many boundary problems (especially in geometries without high degree of symmetry), numerical methods is the only way to the solution. Despite the current abundance of software codes and packages offering their automatic numerical solution,⁶⁹ it is important to an educated physicist to understand “what is under their hood”, at least because most universal programs exhibit mediocre performance in comparison with custom codes written for particular problems, and sometimes do not converge at all, especially for fast-changing (say, exponential) functions.

The simplest of the numerical methods of solution of partial differential equations is the *finite-difference* method⁷⁰ in which the sought function of N scalar arguments $f(r_1, r_2, \dots, r_N)$ is represented by its values in discrete points of a rectangular grid (also called *mesh*) of the corresponding dimensionality (Fig. 31).

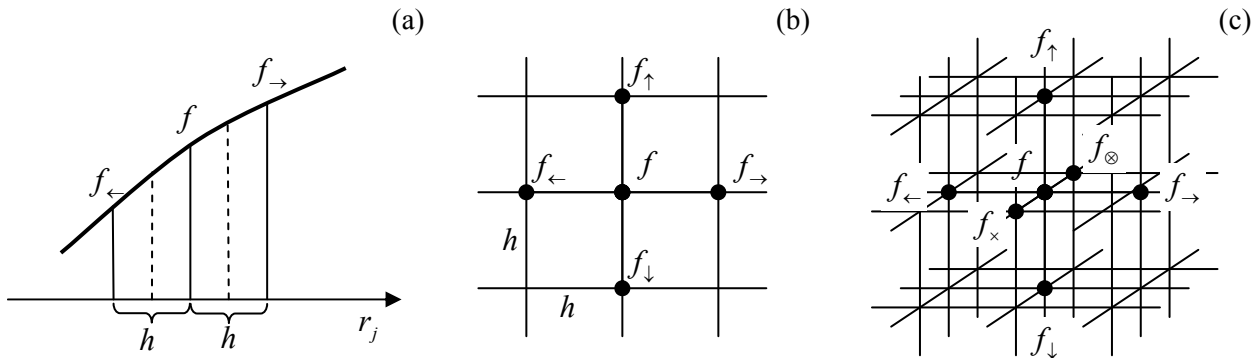


Fig. 2.31. General idea of the finite-difference method in (a) one, (b) two, and (c) three dimensions.

Each partial second derivative of the function is approximated by the formula that readily follows from the linear approximations for the function f and then its partial derivatives – see Fig. 31a:

$$\frac{\partial^2 f}{\partial r_j^2} = \frac{\partial}{\partial r_j} \left(\frac{\partial f}{\partial r_j} \right) \approx \frac{1}{h} \left(\frac{\partial f}{\partial r_j} \Big|_{r_j+h/2} - \frac{\partial f}{\partial r_j} \Big|_{r_j-h/2} \right) \approx \frac{1}{h} \left[\frac{f_{\rightarrow} - f}{h} - \frac{f - f_{\leftarrow}}{h} \right] = \frac{f_{\rightarrow} + f_{\leftarrow} - 2f}{h^2}, \quad (2.220)$$

⁶⁹ See, for example, MA Secs. 16 (iii) and (iv).

⁷⁰ For more details see, e.g., R. Leveque, *Finite Difference Methods for Ordinary and Partial Differential Equations*, SIAM, 2007.

where $f_{\rightarrow} \equiv f(r_j + h)$ and where $f_{\leftarrow} \equiv f(r_j - h)$. (The relative error of this approximation is of the order of $h^4 \partial^4 f / \partial r_j^4$.) As a result, a 2D Laplace operator may be presented as

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = \frac{f_{\rightarrow} + f_{\leftarrow} - 2f}{h^2} + \frac{f_{\uparrow} + f_{\downarrow} - 2f}{h^2} = \frac{f_{\rightarrow} + f_{\leftarrow} + f_{\uparrow} + f_{\downarrow} - 4f}{h^2}, \quad (2.221)$$

while the 3D operator as

$$\frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = \frac{f_{\rightarrow} + f_{\leftarrow} + f_{\uparrow} + f_{\downarrow} + f_{\otimes} + f_{\times} - 6f}{h^2}. \quad (2.222)$$

(The notation used in these formulas should be clear from Figs. 31b and 31c, respectively.)

Let us apply this scheme to find the electrostatic potential distribution inside of a cylindrical box with conducting walls and square cross-section, using an extremely coarse mesh with step $h = a/2$ (Fig. 32). In this case our function, the electrostatic potential, equals zero on the side walls and the bottom, and equals to V_0 at the top lid, so that, according to Eq. (221), the Laplace equation may be approximated as

$$\frac{0 + 0 + V_0 + 0 - 4\phi}{(a/2)^2} = 0. \quad (2.223)$$

The resulting value for the potential in the center of the box is $\phi = V_0/4$. Surprisingly, this is the *exact* value! This may be proved by solving this problem by the variable separation method, just as this has been done for the similar 3D problem in Sec. 4 above. The result is

$$\phi(x, y) = \sum_{n=1}^{\infty} c_n \sin \frac{\pi n x}{a} \sinh \frac{\pi n y}{a}, \quad c_n = \frac{4V_0}{\pi n \sinh(\pi n)} \times \begin{cases} 1, & \text{if } n \text{ is odd,} \\ 0, & \text{otherwise.} \end{cases} \quad (2.224)$$

so that at the central point ($x = y = a/2$),

$$\phi = \frac{4V_0}{\pi} \sum_{j=0}^{\infty} \frac{\sin[\pi(2j+1)/2] \sinh[\pi(2j+1)/2]}{(2j+1) \sinh[\pi(2j+1)]} = \frac{2V_0}{\pi} \sum_{j=0}^{\infty} \frac{(-1)^j}{(2j+1) \cosh[\pi(2j+1)/2]}. \quad (2.225)$$

The last series equals exactly to $\pi/8$, so that $\phi = V_0/4$.

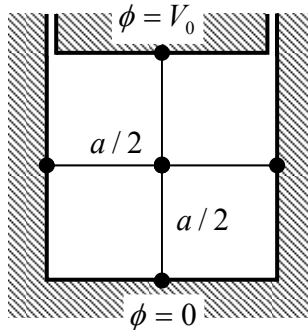


Fig. 2.32. Numerical solution of the internal 2D boundary problem for a conducting, cylindrical box with square cross-section, using a very coarse mesh (with $h = a/2$).

For a similar 3D problem (a cubic box) we can use Eq. (222) to get

$$\frac{0 + 0 + V_0 + 0 + 0 + 0 - 6\phi}{(a/2)^2} = 0, \quad (2.226)$$

so that $\phi = V_0/6$. Unbelievably enough, this result is also exact! (This follows from our variable separation result expressed by Eqs. (95) and (99).)

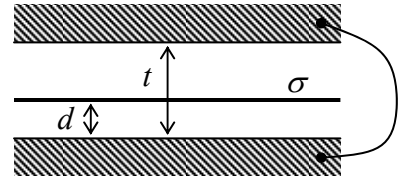
Though such exact results should be considered as a happy coincidence rather than the norm, they still show that numerical methods, with a relatively crude mesh, may be more computationally efficient than the “analytical” approaches, like the variable separation method with its infinite-series results that, in most cases, require computers anyway for the result comprehension and analysis.

A more powerful (but also much more complex for implementation) approach is the *finite-element* method in which the discrete point mesh, typically with triangular cells, is (automatically) generated in accordance with the system geometry. Such mesh generators provide higher point concentration near sharp convex parts of conductor surfaces, where the field concentrates and hence the potential changes faster, and thus ensure better accuracy-to-performance trade-off than the finite-difference methods on a uniform grid. The price to pay for this improvement is the algorithm complexity that makes manual adjustments much harder. Unfortunately I do not have time for going into the details of that method, and have to refer the reader to the special literature on this subject.⁷¹

2.9. Exercise problems

2.1. Calculate the force (per unit area) exerted on a conducting surface by an electric field. Compare the result with the definition of the electric field, given by Eq. (1.5).

2.2. A thin plane film, carrying a uniform electric charge density σ , is placed inside a plane capacitor whose plates are connected by a wire – see Fig. on the right. Neglecting the edge effects, calculate the surface charges of the plates, and the net force acting on the film (per unit area).

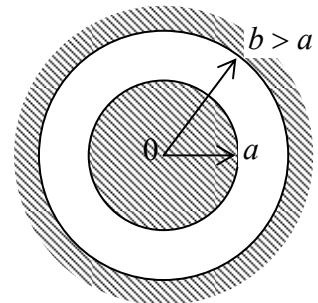


2.3. Following the discussion of two weakly coupled spheres in Sec. 2, find an approximate expression for the mutual capacitance (per unit length) between two very thin, parallel wires, both with a round cross-section, but each with its own diameter. Compare the result with that for two small spheres, and interpret the difference.

2.4. Use the Gauss law to calculate the mutual capacitance of the following 2-electrode systems, with the cross-section shown in Fig. 5 (reproduced on the right):

(i) a conducting sphere inside a concentric spherical cavity in another conductor, and

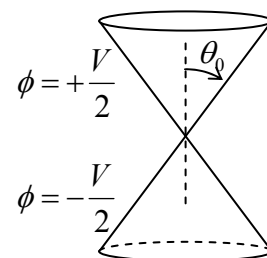
(ii) a conducting cylinder inside a coaxial cavity in another conductor. (In this case, we speak about the capacitance per unit length).



⁷¹ See, e.g., C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Dover, 2009, or T. J. R. Hughes, *The Finite Element Method*, Dover, 2000.

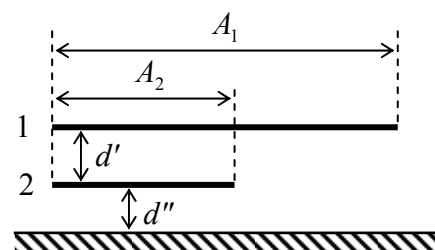
Compare the results with those obtained in Sec. 2.2, using the Laplace equation solution.

2.5. Calculate the electrostatic potential distribution around two barely separated conductors in the form of coaxial, round cones (see Fig. on the right), with voltage V between them. Compare the result with that of a similar 2D problem, with the cones replaced by plane-face wedges. Can you calculate the mutual capacitance between the conductors in any of these systems? If not, can you estimate it?



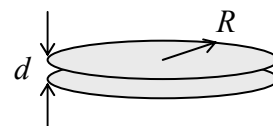
2.6. A system of two thin conducting plates is located over a ground plane as shown in Fig. on the right, where A' and A'' are plate part areas, while d' and d'' are distances between them. Neglecting the fringe effects, calculate:

- (i) the effective capacitance of each plate, and
- (ii) their mutual capacitance.



2.7. Using the results for a single thin round disk, obtained in Sec. 4, consider a system of two such disks at a small distance $d \ll R$ from each other - see Fig. on the right. In particular, calculate:

- (i) the reciprocal capacitance matrix of the system,
- (ii) the mutual capacitance between the disks,
- (iii) the partial capacitance, and
- (iv) the effective capacitance of one disk,



(all in the first non-vanishing approximations in $d/R \ll 1$). Compare the results (ii)-(iv) and interpret their similarities and differences.

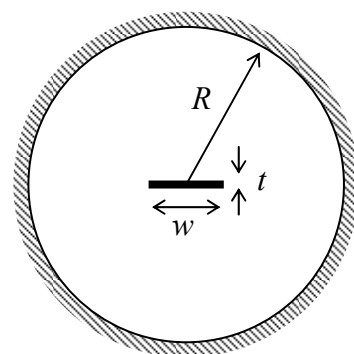
2.8.* Calculate the mutual capacitance (per unit length) between two cylindrical conductors forming a system with the cross-section shown in Fig. on the right, in the limit $t \ll w \ll R$.

Hint: You may like to use *elliptical* (not “ellipsoidal”!) coordinates $\{\alpha, \beta\}$ defined by the following equation:

$$x + iy = c \cosh(\alpha + i\beta), \quad (*)$$

with the appropriate choice of constant c . In these orthogonal 2D coordinates, the Laplace operator is very simple:⁷²

$$\nabla^2 = \frac{1}{c^2 (\cosh^2 \alpha - \cos^2 \beta)} \left(\frac{\partial^2}{\partial \alpha^2} + \frac{\partial^2}{\partial \beta^2} \right).$$



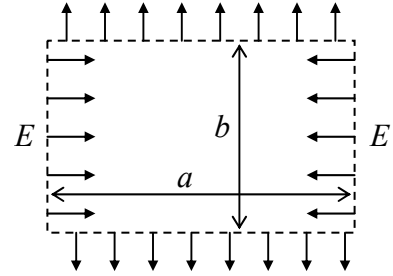
⁷² This fact should not be surprising, because Eq. (*) is essentially the conformal map $z = c \cosh \mathfrak{u}$, where $z = x + iy$, and $\mathfrak{u} = \alpha + i\beta$ - see the discussion in Sec. 4.

2.9. Formulate 2D electrostatic problems that can be solved using each of the following analytic functions of the complex variable $z \equiv x + iy$:

- (i) $w = \ln z$,
- (ii) $w = z^{1/2}$,

and solve these problems.

2.10. On each wall of a cylindrical volume with a rectangular cross-section $a \times b$, with no electric charges inside it, the electric field is uniform, normal to the wall plane, and opposite to that on opposite side – see Fig. on the right. Calculate the distribution of the electric potential inside the volume, provided that the field magnitude on the vertical walls equals E .



2.11. Complete the solution of the problem shown in Fig. 10, by calculating the distribution of the surface charge of the semi-planes. Can you calculate the mutual capacitance between the plates (per unit length)? If not, can you estimate it?

2.12.* A straight, long, thin, round-cylindrical metallic pipe has been cut, along its axis, into two equal parts – see Fig. on the right.

(i) Use the conformal mapping method to calculate the distribution of the electrostatic potential, created by voltage V applied between the two parts, both outside and inside the pipe, and of the surface charge.

(ii) Calculate the mutual capacitance between pipe's halves (per unit length), taking into account a small width $2t \ll R$ of the cut.

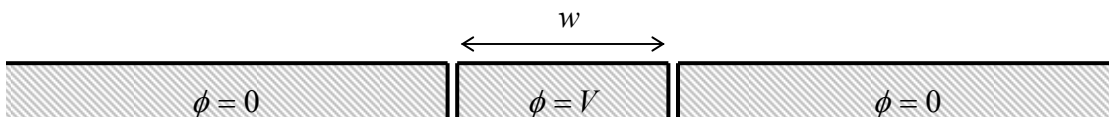
Hints: In Task (i), you may like to use the following complex function:

$$w = \ln \left(\frac{R+z}{R-z} \right),$$

while in Task (ii), it is advisable to use the solution of the previous problem.

2.13. Solve Task (i) of the previous problem using the variable separation method, and compare the results.

2.14. Use the variable separation method to calculate the potential distribution above the plane surface of a conductor, with a strip of width w separated by very thin cuts, and biased with voltage V – see Fig. below.



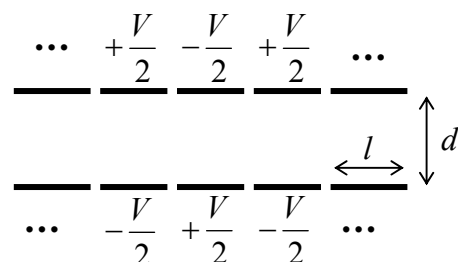
2.15. In the Fig. of the previous problem, the cut-out and voltage-biased part of the conducting plane is now not a strip, but a square with side w . Calculate the potential distribution above conductor's surface.

2.16. Complete the cylinder problem started in Sec. 5 (see Fig. 15), for the cases when voltage on the top lid is fixed as follows:

- (i) $V = V_0 J_1(\xi_{11} \rho / R) \sin \varphi$, where $\xi_{11} \approx 3.832$ is the first root of function $J_1(x)$, and
- (ii) $V = V_0 = \text{const.}$

For both cases, calculate the electric field in the centers of the lower and upper lids. (For assignment (ii), an answer including series and/or integrals is satisfactory.)

2.17. Each electrode of a large plane capacitor is cut into long strips of equal width l , with very narrow gaps between them. These strips are kept at the alternating potentials as shown in Fig. on the right. Use the variable separation method to calculate the electrostatic potential distribution. Explore the limit $l \ll d$.



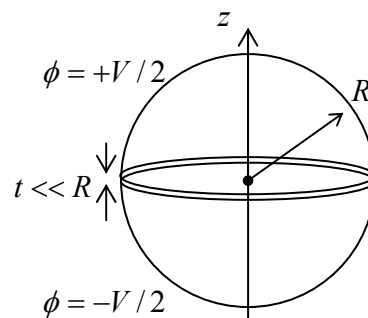
2.18. Solve the problem shown in Fig. 19. In particular:

- (i) calculate and sketch the distribution of the electrostatic potential inside the system for various values of ratio R/h , and
- (ii) simplify the results for the limit $R/h \rightarrow 0$.

2.19. Use the variable separation method to find the potential distribution inside and outside of a thin spherical shell of radius R , with fixed potential $\phi(R, \theta, \varphi) = V_0 \sin \theta \cos \varphi$.

2.20. A thin spherical shell carries charge with areal density $\sigma = \sigma_0 \cos \theta$. Calculate the spatial distribution of the electrostatic potential and field.

2.21. Use the variable separation method to calculate the potential distribution both inside and outside of a thin spherical shell of radius R , separated with a very thin cut, along plane $z = 0$, into two halves, with voltage V applied between them – see Fig. on the right. Analyze the solution; in particular, compare the field at axis z , for $z > R$, with Eq. (2.218), obtained by the Green's function method.

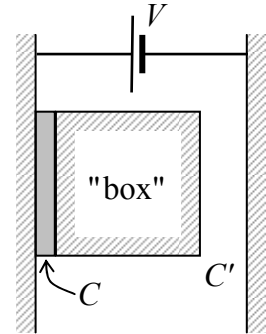


Hint: You may like to use the following integral of a Legendre polynomial with odd index $l = 1, 3, 5 \dots = 2n - 1$:⁷³

⁷³ As a reminder, the *double factorial* (also called “semifactorial”) operator (!) is similar to the usual factorial operator (!), but with the product limited to numbers of the same parity as its argument (in our particular case, of the odd numbers in the nominator, and even numbers in the denominator).

$$I_n \equiv \int_0^1 P_{2n-1}(\xi) d\xi = \frac{1}{n!} \cdot \left(\frac{1}{2}\right) \cdot \left(-\frac{3}{2}\right) \cdot \left(-\frac{5}{2}\right) \cdots \left(\frac{3}{2} - n\right) \equiv (-1)^{n-1} \frac{(2n-3)!!}{2n(2n-2)!!}.$$

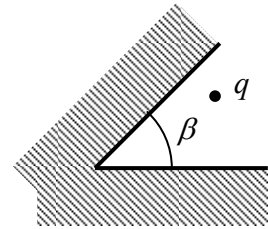
2.22. A small conductor (in this context, usually called the *single-electron box* or *single-electron island*) is placed between two conducting electrodes, with voltage V applied between them. The gap between one of the electrodes and the box is so narrow that electrons may tunnel quantum-mechanically through this gap (“weak tunnel junction”) – see Fig. on the right. Neglecting thermal fluctuations, calculate the equilibrium charge of the island as a function of V .



Hint: To solve this problem, you do not need to know much about quantum-mechanical tunneling through weak junctions,⁷⁴ besides that such tunneling of an electron, and its subsequent energy relaxation inside the conductor, may be considered as a single inelastic (energy-dissipating) event. In the absence of thermal agitation, such event takes place when (and only when) it decreases the potential energy of the system.

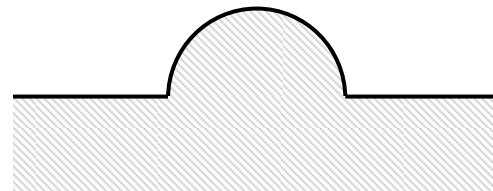
2.23. Use the image charge method to calculate the surface charges induced in the plates of a very broad plane capacitor of thickness D by a point charge q separated from one of the electrodes by distance d .

2.24. Prove the statement, made in Sec. 6, that the 2D boundary problem shown in Fig. on the right can be solved using a finite number of image charges if angle β equals π/n , where $n = 1, 2, \dots$



2.25. Use the image charge method to calculate the energy of electrostatic interaction between a point charge placed in the center of a spherical cavity that was carved inside a grounded conductor, and the conductor’s walls. Looking at the result, could it be obtained in a simpler way (or ways)?

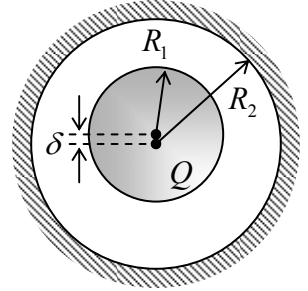
2.26. Use the method of images to find the Green’s function of the system shown in Fig. on the right, where the bulge on the conducting plane has the shape of a semi-sphere of radius R .



2.27.* Use the fact of spherical inversion, expressed by Eq. (198), to develop an iterative method for more and more precise calculation of the mutual capacitance between two similar metallic spheres of radius R , with centers separated by distance $d > 2R$.

⁷⁴ In this context, weak junction means a tunnel junction with transparency so low that the tunneling electron’s wavefunction loses its quantum-mechanical coherence before the electron has time to tunnel back. In a typical junction of a macroscopic area this condition is fulfilled if the effective tunnel resistance of the junction is much higher than the quantum unit of resistance (see, e.g., QM Sec. 3.2), $R_Q \equiv \pi\hbar/2e^2 \approx 6.5 \text{ k}\Omega$.

2.28.* A metallic sphere of radius R_1 , carrying electric charge Q , is placed inside a spherical cavity of radius $R_2 > R_1$, cut inside another metal. Calculate the force exerted on the sphere if its center is displaced by a small distance $\delta \ll R_1, R_2 - R_1$ from that of the cavity – see Fig. on the right.



2.29. Within the simple model of electric field screening in conductors, discussed in Sec. 2.1, analyze the partial screening of the electric field of a point charge q by a plane, uniform conducting film of thickness $t \ll \lambda$, where λ is (depending on charge carrier statistics) either the Debye or the Thomas-Fermi screening length – see, respectively, Eqs. (2.8) or (2.10). Assume that the distance d between the charge and plane is much larger than t .

2.30. Suggest a convenient definition of 2D Green's function for electrostatic problems, and calculate it for:

- (i) the unlimited free space, and
- (ii) the free space above a conducting plane.

Use the latter result to re-solve Problem 14.

2.31. Find the 2D Green's function for the free space

- (i) outside a round conducting cylinder,
- (ii) inside a round cylindrical hole in a conductor.

2.32. Solve Task (i) of Problem 12 (see also Problem 13), using the Green's function method.

Hints: You may like to use the 2D Green's function derived in the solution of Problem 2.27(ii), and the following table integral:⁷⁵

$$\int \frac{d\xi}{a + b \cos \xi} = \frac{2}{(a^2 - b^2)^{1/2}} \tan^{-1} \left[\frac{(a - b)}{(a^2 - b^2)^{1/2}} \tan \frac{\xi}{2} \right], \quad \text{if } a^2 - b^2 > 0.$$

2.33. Solve the same 2D boundary problem that was discussed in Sec. 6 (Fig. 32) using:

- (i) the finite difference method, with a finer square mesh, $h = a/3$, and
- (ii) the variable separation method.

Compare the results (at the mesh points only) and comment.

⁷⁵ Here the notation \tan^{-1} is used for the multi-valued function (alternatively called Arctan) which is reciprocal to \tan . (Due to the π -periodicity of the \tan , function \tan^{-1} is defined to an arbitrary additive multiple of π .) At the value interval $[-\pi/2, +\pi/2]$, \tan^{-1} is usually called \arctan .

This page is
intentionally left
blank

Chapter 3. Polarization of Dielectrics

In the last chapter, we have discussed the electric polarization of conductors. In contrast to those materials, in dielectrics the charge motion is limited to the interior of an atom or a molecule, so that the electric polarization of these materials by external field takes a different form. This issue is the main subject of this chapter. In preparation to the analysis of dielectrics, we have to start with a more general discussion of the electric field of a spatially-restricted system of charges.

3.1. Electric dipole

Let us consider a localized system of charges, of a linear size scale a , and calculate a simple but approximate expression for the electrostatic field induced by the system at a distant point \mathbf{r} . For that, let us select a reference frame with the origin either somewhere inside the system, or at a distance of the order of a from it (Fig. 1).

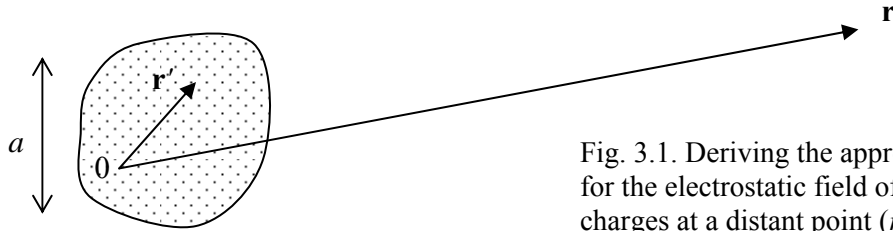


Fig. 3.1. Deriving the approximate expression (5) for the electrostatic field of a localized system of charges at a distant point ($r \gg r' \sim a$).

Then positions of all charges of the system satisfy the following condition

$$r' \ll r. \quad (3.1)$$

Using this condition, we can expand the general expression (1.38) for the electrostatic potential $\phi(\mathbf{r})$ of the system into the Taylor series in small parameter $\mathbf{r}' \equiv \{r'_1, r'_2, r'_3\}$. For any spatial function of the type $f(\mathbf{r} - \mathbf{r}')$, the expansion may be presented as¹

$$f(\mathbf{r} - \mathbf{r}') \approx f(\mathbf{r}) - \sum_{j=1}^3 r'_j \frac{\partial f}{\partial r_j}(\mathbf{r}) + \frac{1}{2!} \sum_{j,j'=1}^3 r'_j r'_{j'} \frac{\partial^2 f}{\partial r_j \partial r_{j'}}(\mathbf{r}) - \dots \quad (3.2)$$

The two leading terms of this expansion, sufficient for our current purposes, may be rewritten in the vector form:²

$$f(\mathbf{r} - \mathbf{r}') \approx f(\mathbf{r}) - \mathbf{r}' \cdot \nabla f(\mathbf{r}) + \dots \quad (3.3)$$

Let us apply this approximate formula to the free-space Green's function (2.204), which weighs the charge density contributions in Eq. (1.38). The gradient of such a spherically-symmetric function $f(r) = 1/r$ is just $\mathbf{n}_r df/dr$, so that

¹ See, e.g., MA Eq. (2.11b).

² The third term (responsible for *quadrupole* effects), as well as all the following, *multipole* terms would require a tensor (rather than vector) representation.

$$\frac{1}{|\mathbf{r} - \mathbf{r}'|} \approx \frac{1}{r} - \mathbf{r}' \cdot \mathbf{n}_r \frac{d}{dr} \left(\frac{1}{r} \right) = \frac{1}{r} + \mathbf{r}' \cdot \frac{\mathbf{r}}{r^3}. \quad (3.4)$$

Plugging this *dipole expansion* into Eq. (1.38), we get

$$\phi(\mathbf{r}) \approx \frac{1}{4\pi\epsilon_0} \left[\frac{1}{r} \int \rho(\mathbf{r}') d^3 r' + \frac{\mathbf{r}}{r^3} \cdot \int \rho(\mathbf{r}') \mathbf{r}' d^3 r' \right] = \frac{1}{4\pi\epsilon_0} \left(\frac{Q}{r} + \frac{\mathbf{r} \cdot \mathbf{p}}{r^3} \right), \quad (3.5)$$

where Q is the net charge of the system, while the vector

$$\mathbf{p} \equiv \int \rho(\mathbf{r}') \mathbf{r}' d^3 r', \quad (3.6)$$

Electric
dipole
moment

with magnitude p of the order of Qa , is called its (electric) *dipole moment*.³

If $Q \neq 0$, the second term in the right-hand part of Eq. (5) is just a small correction to the first one, and in many cases may be ignored. (Remember, Eq. (5) is only valid in the limit $r/a \rightarrow \infty$). However, the net charge of many systems is exactly zero. The most important example is a neutral atom or a neutral molecule, in which the negative charge of electrons exactly compensates the positive charge of protons in nuclei. For such neural systems, the second (dipole-moment) term, ϕ_d , in Eq. (5) is the leading one. Due to its importance, let us rewrite this expression in two other, equivalent forms:

$$\phi_d \equiv \frac{1}{4\pi\epsilon_0} \frac{\mathbf{r} \cdot \mathbf{p}}{r^3} = \frac{1}{4\pi\epsilon_0} \frac{p \cos \theta}{r^2} = \frac{1}{4\pi\epsilon_0} \frac{pz}{[x^2 + y^2 + z^2]^{3/2}}, \quad (3.7)$$

Electric
dipole's
potential

that are more convenient for some applications. Here θ is the angle between vectors \mathbf{p} and \mathbf{r} , and in the last (Cartesian) presentation, axis z is directed along vector \mathbf{p} . Figure 2a shows equipotential surfaces of the dipole field (or rather their cross-sections by a plane in which vector \mathbf{p} resides).

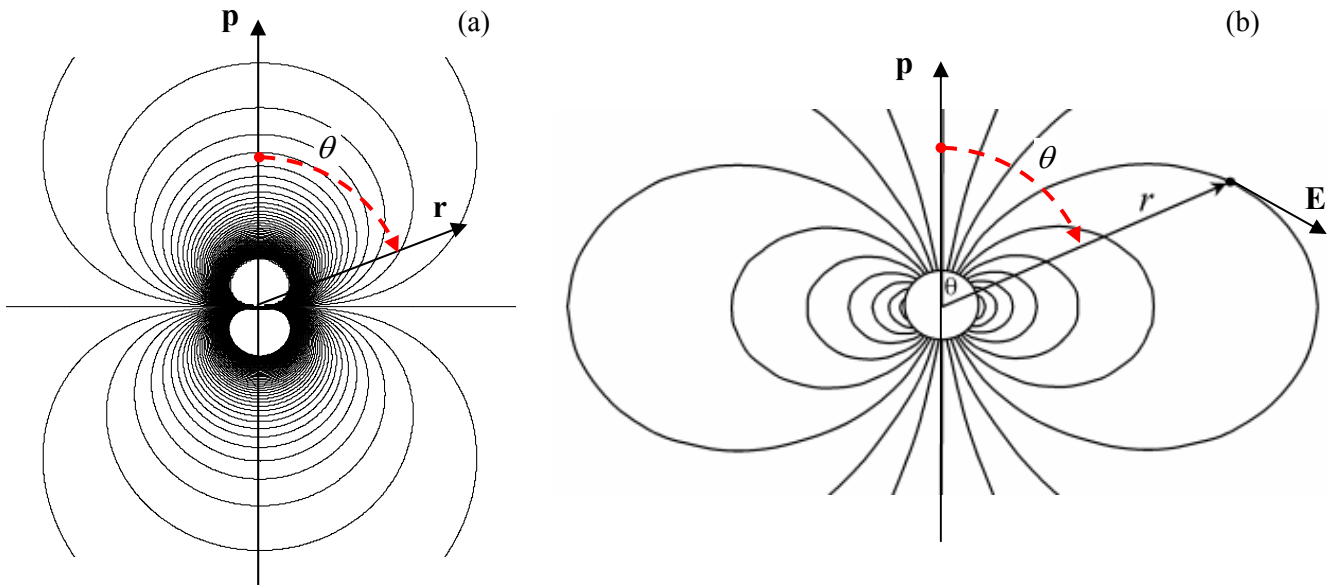


Fig. 3.2. Dipole field: (a) equipotential surfaces and (b) electric field lines, for vertical vector \mathbf{p} .

³ Accordingly, a localized system of charges with $Q = 0$, but $\mathbf{p} \neq 0$, is called an (electric) *dipole*.

The simplest example of the dipole (that gave such systems their name) is a system of two equal but opposite point charges, $+q$ and $-q$, with radius-vectors, respectively, \mathbf{r}_+ and \mathbf{r}_- :

$$\rho(\mathbf{r}) = (+q)\delta(\mathbf{r} - \mathbf{r}_+) + (-q)\delta(\mathbf{r} - \mathbf{r}_-). \quad (3.8)$$

For this system, Eq. (6) yields

$$\mathbf{p} = (+q)\mathbf{r}_+ + (-q)\mathbf{r}_- = q(\mathbf{r}_+ - \mathbf{r}_-) = q\mathbf{a}, \quad (3.9)$$

where \mathbf{a} is the vector connecting points \mathbf{r}_- and \mathbf{r}_+ . Note that in this case (and for all systems with $Q = 0$), the dipole moment does not depend on the reference frame origin choice.

A less trivial example is a conducting sphere of radius R in a uniform external electric field \mathbf{E}_0 . As a reminder, we have solved this problem in Sec. 2.5(iv) and obtained Eq. (2.176) as a result. The first term in the parentheses of that relation describes the external field (2.173), so that the field of the sphere itself (meaning the field of its surface charge induced by \mathbf{E}_0) is given by the second term:

$$\phi_s = \frac{E_0 R^3}{r^2} \cos \theta. \quad (3.10)$$

Comparing this expression with the second form of Eq. (7), we see that the sphere has an *induced* dipole moment

$$\mathbf{p} = 4\pi\epsilon_0 \mathbf{E}_0 R^3. \quad (3.11)$$

This is an interesting example of a *purely* dipole field – in all points outside the sphere ($r > R$), the field has no higher moments.⁴

Returning to the general properties of the dipole field, let us calculate its characteristics. First of all, we may use Eq. (7) to calculate the electric field of a dipole:

$$\mathbf{E}_d = -\nabla\phi_d = -\frac{1}{4\pi\epsilon_0} \nabla \left(\frac{\mathbf{r} \cdot \mathbf{p}}{r^3} \right) = -\frac{1}{4\pi\epsilon_0} \nabla \left(\frac{p \cos \theta}{r^2} \right). \quad (3.12)$$

The differentiation is easiest in spherical coordinates, using the following well-known expression for the gradient of a scalar function in these coordinates⁵ and taking axis z parallel to vector \mathbf{p} . From the last form of Eq. (12) we immediately get

Electric
dipole's
field

$$\mathbf{E}_d = \frac{p}{4\pi\epsilon_0 r^3} (2\mathbf{n}_r \cos \theta + \mathbf{n}_\theta \sin \theta) = \frac{1}{4\pi\epsilon_0} \frac{3\mathbf{r}(\mathbf{r} \cdot \mathbf{p}) - \mathbf{p}r^2}{r^5}. \quad (3.13)$$

Figure 2b shows the electric field lines given by Eqs. (13).

Next, let us calculate the potential energy of interaction between a fixed dipole and a external electric field, using Eq. (1.54). Assuming that the external field does not change much at distances of the order of a (Fig. 1), we may expand the external potential $\phi_{\text{ext}}(\mathbf{r})$ into the Taylor series, just as Eq. (3) prescribes, and keep only its two leading terms:

⁴ Other examples of dipole fields are given by two more systems discussed in Chapter 2 – see Eqs. (2.215) and (2.219). Those systems, however, do have higher-order multipole moments, so that for them, Eq. (7) gives only the long-distance approximation.

⁵ See, e.g., MA Eq. (10.8) with $\partial/\partial\varphi = 0$.

$$U = \int \rho(\mathbf{r}) \phi_{\text{ext}}(\mathbf{r}) d^3r \approx \int \rho(\mathbf{r}) [\phi_{\text{ext}}(0) + \mathbf{r} \cdot \nabla \phi_{\text{ext}}(0)] d^3r = Q\phi_{\text{ext}}(0) - \mathbf{p} \cdot \mathbf{E}_{\text{ext}}. \quad (3.14)$$

The first term is the potential energy the system would have if it were a point charge. If the net charge Q is zero, that term disappears, and the leading contribution is due to the dipole moment:

$$U = -\mathbf{p} \cdot \mathbf{E}_{\text{ext}}. \quad (3.15)$$

Dipole's
energy in
external
field

Note, however, that Eq. (15) is only valid for a fixed dipole, with \mathbf{p} independent of \mathbf{E}_{ext} . In the opposite limit, when the dipole is *induced* by the field, i.e. $\mathbf{p} \propto \mathbf{E}_{\text{ext}}$ (see Eq. (11) as an example), we can repeat the discussion that accompanied Fig. 1.6 to show that Eq. (15) acquires an additional factor $1/2$.

In particular, combining Eqs. (13) and Eq. (15), we may get the following important formula for interaction of two independent dipoles

$$U = \frac{1}{4\pi\epsilon_0} \frac{\mathbf{p}_1 \cdot \mathbf{p}_2 r^2 - 3(\mathbf{r} \cdot \mathbf{p}_1)(\mathbf{r} \cdot \mathbf{p}_2)}{r^5} = \frac{1}{4\pi\epsilon_0} \frac{p_{1x}p_{2x} + p_{1y}p_{2y} - 2p_{1z}p_{2z}}{r^3}, \quad (3.16)$$

where \mathbf{r} is the vector connecting the dipoles, and axis z is directed along this vector. If each moment is due to the polarization of the dipole by the electric field of its counterpart: $\mathbf{p}_{1,2} \propto \mathbf{E}_{2,1} \propto 1/r^3$, this expression (which is valid for this case with the additional factor $1/2$) the potential is always negative and proportional to $1/r^6$. Such potential describes, in particular, the long-range, attractive part (the so-called *London dispersion force*) of the interaction between electrically neutral atoms and molecules.⁶

According to Eq. (15), in order to reach the minimum of U , the electric field “tries” to align the dipole direction along its own. The quantitative expression of this effect is the torque $\boldsymbol{\tau}$ exerted by the field. The simplest way to calculate it is to sum up all the elementary torques $d\boldsymbol{\tau} = \mathbf{r} \times d\mathbf{F}_{\text{ext}} = \mathbf{r} \times \mathbf{E}_{\text{ext}}(\mathbf{r})\rho(\mathbf{r})d^3r$ exerted on all elementary charges of the system:

$$\boldsymbol{\tau} = \int \mathbf{r} \times \mathbf{E}_{\text{ext}}(\mathbf{r})\rho(\mathbf{r})d^3r \approx \mathbf{p} \times \mathbf{E}_{\text{ext}}(0), \quad (3.17)$$

where at the last transition we have again neglected the spatial dependence of the external field.

The spatial dependence of \mathbf{E}_{ext} cannot, however, be ignored at the calculation of the *total force* exerted by the field on the dipole (with $Q = 0$). Indeed, Eq. (15) shows that if the field is constant, the dipole energy is the same at all spatial points, and hence the net force is zero. However, if the field has a finite gradient, a total force does appear:

$$\mathbf{F} = -\nabla U = \nabla(\mathbf{p} \cdot \mathbf{E}_{\text{ext}}), \quad (3.18)$$

where the derivative has to be taken at the dipole's position (in our notation, at $\mathbf{r} = 0$). If the dipole that is being moved in a field retains its magnitude and orientation, then the last formula is equivalent to⁷

$$\mathbf{F} = (\mathbf{p} \cdot \nabla) \mathbf{E}_{\text{ext}}. \quad (3.19)$$

Alternatively, the last expression may be obtained similarly to Eq. (14):

$$\mathbf{F} = \int \rho(\mathbf{r}) \mathbf{E}_{\text{ext}}(\mathbf{r}) d^3r \approx \int \rho(\mathbf{r}) [\mathbf{E}_{\text{ext}}(0) + (\mathbf{r} \cdot \nabla) \mathbf{E}_{\text{ext}}] d^3r = Q\mathbf{E}_{\text{ext}}(0) + (\mathbf{p} \cdot \nabla) \mathbf{E}_{\text{ext}}. \quad (3.20)$$

⁶ See, e.g., SM Sec. 3.5.

⁷ The equivalence may be proved, for example, by using MA Eq. (11.6) with $\mathbf{f} = \mathbf{p} = \text{const}$ and $\mathbf{g} = \mathbf{E}_{\text{ext}}$, taking into account that according to the general Eq. (1.28), $\nabla \times \mathbf{E}_{\text{ext}} = 0$.

Finally, let me add a note on the so-called *coarse-grain model* of the dipole. The dipole approximation explored above is asymptotically correct *at large distances*, $r \gg a$. However, for some applications (including the forthcoming discussion in Sec. 5 of *molecular field effects*) it is important to have an expression that would be approximately valid *everywhere* in space, though maybe without exact details at $r \sim a$, and also give the correct result for the space-average of the electric field,

$$\bar{\mathbf{E}} \equiv \frac{1}{V} \int_V \mathbf{E} d^3r, \quad (3.21)$$

where V is a regularly-shaped volume much larger than a^3 , for example a sphere of radius $R \gg a$, with the dipole at its center. For the field \mathbf{E}_d given by Eq. (13), such average is zero. Indeed, let us consider Cartesian components of that vector in the coordinate system with axis z directed along vector \mathbf{p} . Due to the axial symmetry of the field, the averages of components E_x and E_y evidently vanish. Let us use Eq. (13) to spell out the “vertical” component of the field (parallel to the dipole moment vector):

$$E_z \equiv \mathbf{E}_d \cdot \frac{\mathbf{p}}{p} = \frac{1}{4\pi\epsilon_0 r^3} (2\mathbf{n}_r \cdot \mathbf{p} \cos\theta - \mathbf{n}_\theta \cdot \mathbf{p} \sin\theta) = \frac{p}{4\pi\epsilon_0 r^3} (2\cos^2\theta - \sin^2\theta). \quad (3.22)$$

Integrating this expression over the whole solid angle $\Omega = 4\pi$, at fixed r , using a convenient variable substitution $\cos\theta \equiv \xi$, we get

$$\oint_{4\pi} E_z d\Omega = 2\pi \int_0^\pi E_z \sin\theta d\theta = \frac{p}{2\epsilon_0 r^3} \int_0^\pi (2\cos^2\theta - \sin^2\theta) \sin\theta d\theta = \frac{p}{2\epsilon_0 r^3} \int_{-1}^{+1} (3\xi^3 - \xi) d\xi = 0. \quad (3.23)$$

On the other hand, the *exact* electric field of an *arbitrary* charge distribution, having the total dipole moment \mathbf{p} , satisfies the following condition,

$$\int_V \mathbf{E}(\mathbf{r}) d^3r = -\frac{\mathbf{p}}{3\epsilon_0}, \quad (3.24)$$

where the integration is over any sphere containing all the charges. A proof of this formula for the general case requires a somewhat cumbersome, though straightforward integration,⁸ but later in the course we will see that it is correct for several particular cases. The origin of the difference between Eqs. (23) and (24) is illustrated in Fig. 3 on the example of a dipole created by two equal but opposite charges – see Eqs. (8)-(9). The zero average of the dipole field (13) does not take into account the contribution of the field in the region between the charges (where Eq. (13) is not valid), which is directed mostly against the dipole vector (9).

Thus in order to be used as a reasonable coarse-grain model, Eq. (13) should be modified as follows:

$$\mathbf{E}_{cg} = \frac{1}{4\pi\epsilon_0} \left[\frac{3\mathbf{r}(\mathbf{r} \cdot \mathbf{p}) - \mathbf{p}r^2}{r^5} - \frac{4\pi}{3} \mathbf{p} \delta(\mathbf{r}) \right]. \quad (3.25)$$

Evidently, such modification does not change the field at large distances $r \gg a$, i.e. in the region where the expansion (3) and hence Eq. (13) are valid.

⁸ See, e.g., the end of Sec. 4.1 in the textbook by J. Jackson, *Classical Electrodynamics*, 3rd ed., Wiley, 1999.

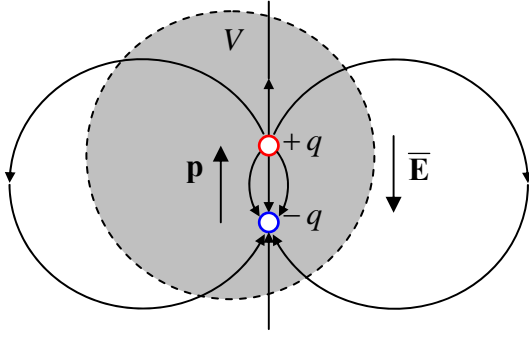


Fig. 3.3. Illustrating the origin of Eq. (24).
(The field lines are *very* approximate.)

3.2. Dipole media

Let us generalize equation (7) to the case of several (possibly, many) dipoles \mathbf{p}_j located at arbitrary points \mathbf{r}_j . Using the linear superposition principle, we get

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \sum_j \mathbf{p}_j \cdot \frac{\mathbf{r} - \mathbf{r}_j}{|\mathbf{r} - \mathbf{r}_j|^3}. \quad (3.26)$$

If our system (medium) contains many similar dipoles, distributed in space with density $n(\mathbf{r})$, we may use the same standard argumentation that has led us from Eq. (1.5) to Eq. (1.8), to rewrite the last sum as an integral

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int \mathbf{P}(\mathbf{r}') \cdot \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} d^3 r', \quad (3.27)$$

where vector $\mathbf{P}(\mathbf{r}) \equiv n(\mathbf{r})\mathbf{p}$, called *electric polarization* has the physical meaning of the net dipole moment per unit volume. Note again that since Eq. (26) does not describe that field at distances comparable to the dipole size, and hence Eq. (27), and all the following formulas of this section, describes the so-called *macroscopic electric field*, i.e. the dipole field averaged over the *microscopic* (dipole-size) distances.

Now comes a very impressive mathematical trick. Just as has been done in the previous section (just with the appropriate sign change), Eq. (27) may be rewritten in the equivalent form

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int \mathbf{P}(\mathbf{r}') \cdot \nabla' \frac{1}{|\mathbf{r} - \mathbf{r}'|} d^3 r', \quad (3.28)$$

where ∇' means the del operator (in this particular case, the gradient) acting in the “source space” of vectors \mathbf{r}' . The right-hand part of Eq. (28), applied to any volume V limited by surface S , may be integrated by parts in the following way:⁹

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \oint_S P_n(\mathbf{r}') \frac{1}{|\mathbf{r} - \mathbf{r}'|} d^2 r' - \frac{1}{4\pi\epsilon_0} \int_V \frac{\nabla' \cdot \mathbf{P}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3 r'. \quad (3.29)$$

⁹ To prove this (almost evident) formula strictly, it is sufficient to apply the divergence theorem given by MA Eq. (12.2), to vector function $\mathbf{f} = \mathbf{P}(\mathbf{r}')/|\mathbf{r} - \mathbf{r}'|$, in the “source space” of radius-vectors \mathbf{r}' .

If the surface does not carry an infinitely dense (δ -functional) sheet of additional dipoles, or it is just very far, the first term in the right-hand part is negligible. Now comparing the second term with the basic equation (1.38) for the electric potential, we see that this term may be interpreted as the field of certain *effective* electric charges with density

$$\rho_{\text{ef}} = -\nabla \cdot \mathbf{P}. \quad (3.30)$$

Effective
charge
density

Figure 4 illustrates the physics of this relation for a cartoon model of a multi-dipole system: a layer of uniformly-distributed two-point-charge units oriented perpendicular to the layer surface. (In this case $\nabla \cdot \mathbf{P} = dP/dx$.) One can see that ρ_{ef} , defined by Eq. (30), may be interpreted as the density of uncompensated surface charges of polarized elementary dipoles.

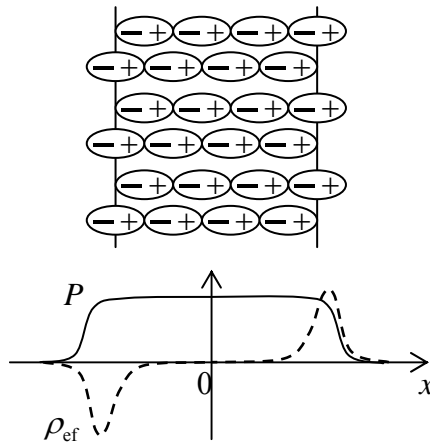


Fig. 3.4. Spatial distributions of the polarization and effective charges in a layer of similar elementary dipoles (schematically).

Next, from Sec. 1.2, we already know that Eq. (1.38) is equivalent to the inhomogeneous Maxwell equation (1.27) for the electric field. This is why Eq. (30) implies that if, besides the compensated charges of the dipoles, the system also has certain *stand-alone charges* (not a part of the dipoles!) distributed in space with density $\rho(\mathbf{r})$, the average electric field obeys, instead of Eq. (1.27), the following generalized equation

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0} (\rho + \rho_{\text{ef}}) = \frac{1}{\epsilon_0} (\rho - \nabla \cdot \mathbf{P}). \quad (3.31)$$

It is evidently tempting (and very convenient for applications!) to carry over the dipole-related term of this equation over to the left-hand part of Eq. (31), and rewrite the resulting equality as the so-called *macroscopic Maxwell equation*

Electric
displacement

$$\nabla \cdot \mathbf{D} = \rho, \quad (3.32)$$

where a new vector, called the *electric displacement*, is defined as¹⁰

¹⁰ Note that the dimensionality of \mathbf{D} in SI units is different from that of \mathbf{E} . In contrast, in the Gaussian units the electric displacement is defined as $\mathbf{D} = \mathbf{E} + 4\pi\mathbf{P}$, so that $\nabla \cdot \mathbf{D} = 4\pi\rho$ (the relation $\rho_{\text{ef}} = -\nabla \cdot \mathbf{P}$ remains the same as in SI units), and the dimensionalities of \mathbf{D} and \mathbf{E} coincide. Philosophically, this coincidence is a certain handicap, because it is frequently convenient to consider Cartesian components of \mathbf{E} as a generalized force, and those of \mathbf{D} as a generalized coordinate (see Sec. 6 below), and it is comforting to have their dimensionality different.

$$\mathbf{D} \equiv \epsilon_0 \mathbf{E} + \mathbf{P}. \quad (3.33)$$

The comparison of Eqs. (32) and (1.27) shows that \mathbf{D} may be interpreted as the “would-be” electric field that *would be* created by stand-alone charges in the absence of the dipole medium polarization. In contrast, \mathbf{E} is the *actual* electric field - though, as was mentioned above, space-averaged over a volume much larger than that of an elementary dipole.¹¹

To complete the general analysis of the multi-dipole systems, let us rewrite the macroscopic Maxwell equation (32) in the integral form. Applying the divergence theorem to an arbitrary volume V limited by surface S , we get the following *macroscopic Gauss law*:

$$\oint_S D_n d^2 r = \int_V \rho d^3 r \equiv Q, \quad (3.34) \quad \text{Macroscopic Gauss law}$$

where Q is the total stand-alone charge inside volume V .

Let me emphasize again that the key Eq. (27), and hence all the following equations of this section, only to the *macroscopic* field, i.e. the electric field averaged over its rapid variations at the atomic space scale. Such macroscopic description is valid as soon as we are not concerned with the inter-atomic field variations - for whose description the classical physics is inadequate in any case.

3.3. Linear dielectrics

The general equations derived above are broadly used to describe any dielectrics – materials with bound electric charges (and hence with no dc electric conduction). The polarization properties of these materials may be described by the dependence between vectors \mathbf{P} and \mathbf{E} . In the most materials, in the absence of external electric field, the elementary dipoles \mathbf{p} either equal zero or have a random orientation in space, so that the net dipole moment of each macroscopic volume (still containing many such dipoles) equals zero: $\mathbf{P} = 0$.

Moreover, if the field changes are sufficiently slow, most materials may be characterized by a unique dependence of \mathbf{P} on \mathbf{E} . Then using the Taylor expansion of function $\mathbf{P}(\mathbf{E})$, we may argue that in relatively low electric fields the function should be well approximated by a linear dependence between these two vectors. In an isotropic media, the coefficient of proportionality should be just a scalar.¹² In SI units, this constant is defined by the following relation:

$$\mathbf{P} = \chi_e \epsilon_0 \mathbf{E}, \quad (3.35) \quad \text{Electric susceptibility definition}$$

with the dimensionless constant χ_e called the *electric susceptibility*. However, it is much more common to use, instead of χ_e , another parameter,

$$\epsilon_r \equiv 1 + \chi_e, \quad (3.36) \quad \text{Dielectric constant}$$

¹¹ Note, however, that such averaging does *not* include the inner-dipole fields which is (approximately) described by the second term of Eq. (25).

¹² In anisotropic materials, such as crystals, a susceptibility *tensor* may be necessary to give an adequate description of the linear relation of vectors \mathbf{P} and \mathbf{E} . Fortunately, in most important crystals (such as silicon) the anisotropy of polarization is small, so that they may be reasonably well characterized by scalar susceptibility.

which is sometimes called the *relative electric permittivity*, but much more often, the *dielectric constant*.¹³ This parameter is very convenient, because combining Eqs. (35) and (36),

$$\mathbf{P} = (\epsilon_r - 1)\epsilon_0 \mathbf{E}. \quad (3.37)$$

and then plugging the resulting relation into the general Eq. (33), we get simply¹⁴

$$\mathbf{D} = \epsilon \mathbf{E}, \quad \text{with } \epsilon \equiv \epsilon_0 \epsilon_r = \epsilon_0 (1 + \chi_e). \quad (3.38)$$

where ϵ is called the *electric permittivity* of the material. Table 1 gives values of the dielectric constant for several representative materials.

Table 3.1. Dielectric constants of a few representative (and/or practically important) dielectrics

Material	ϵ_r
Air (at ambient conditions)	1.00054
Teflon (polytetrafluoroethylene, C_nF_{2n})	2.1
Silicon dioxide (amorphous)	3.9
Glasses (of various compositions)	3.7-10
Castor oil	4.5
Silicon	11.7
Water (at 100°C)	55.3
Water (at 20°C)	80.1
Barium titanate ($BaTiO_3$ at 20°C)	~1,600

In order to get some feeling of the physics behind these values, let us consider a very common model of a media whose elementary dipoles do not interact, so that in the relation $\mathbf{P} = n\mathbf{p}$ the elementary dipole moments \mathbf{p} may be calculated independently of each other. This means that in a linear dielectric, in which Eq. (35) holds, each induced dipole moment \mathbf{p} has to be proportional to the applied field \mathbf{E} as well. Let us write this dependence in the following traditional form,

$$\mathbf{p} = 4\pi\epsilon_0\alpha_{\text{mol}}\mathbf{E}, \quad (3.39)$$

where α_{mol} is called the *molecular* (or, sometimes, “atomic”) *polarizability*, so that

$$\mathbf{P} = n\mathbf{p} = 4\pi\epsilon_0\alpha_{\text{mol}}n\mathbf{E}. \quad (3.40)$$

Comparing this relation with Eq. (35), we get $\chi_e = 4\pi\alpha_{\text{mol}}n$, so that Eq. (36) yields¹⁵

¹³ Note that in electrical engineering literature, the dielectric constant is often denoted by letters κ or K .

¹⁴ In Gaussian units, χ_e is defined by relation $\mathbf{P} = \chi_e \mathbf{E}$, while ϵ is still defined as $\mathbf{D} = \epsilon \mathbf{E}$. Because of that, ϵ is dimensionless and equals $(1 + 4\pi\chi_e)$. Note that $(\epsilon)_{\text{Gaussian}} = (\epsilon/\epsilon_0)_{\text{SI}} = \epsilon_r$, and $(\chi_e)_{\text{SI}} = 4\pi(\chi_e)_{\text{Gaussian}}$, sometimes creating a confusion with the numerical values of the latter parameter.

$$\varepsilon_r = 1 + 4\pi\alpha_{\text{mol}}n. \quad (3.41)$$

Now let us consider the following toy model of a dielectric:¹⁶ a set of similar conducting spheres of radius R , spread apart with small density $n \ll 1/R^3$. At such low density of the spheres, their electrostatic interaction is negligible, and we can use Eq. (11) for the induced dipole moment of a single sphere. Then the polarizability definition (39) yields $\alpha_{\text{mol}} = R^3$, so that $\chi_e = 4\pi nR^3$, and

$$\varepsilon_r = 1 + 4\pi R^3 n. \quad (3.42)$$

Let us use this result for a crude estimate of the dielectric constant of air at the so-called *ambient conditions*, meaning the normal atmospheric pressure, and temperature $T = 300$ K. At these conditions the molecular density n may be, with a few-percent accuracy, found from the equation of state of an ideal gas:¹⁷ $n \approx P/k_B T \approx (1.013 \times 10^5)/(1.38 \times 10^{-23} \times 300) \approx 2.5 \times 10^{25} \text{ m}^{-3}$. The main component of air, molecular nitrogen N_2 , has a van-der-Waals radius¹⁸ of $155 \text{ pm} = 1.55 \times 10^{-10} \text{ m}$. Using it for R , from our crude model we get $\varepsilon_r \approx 1.001$. Comparing this number with the first line of Table 1, we see that our crude model gives surprisingly reasonable results: in order to get the exact experimental value, it is sufficient to decrease R by just $\sim 25\%$.

This result may encourage us to try using Eq. (42) for larger density n , i.e., beyond the range of its quantitative applicability. For example, as a crude model for solid and liquids let us assume that spheres form a simple cubic lattice with period $a = 2R$ (i.e., the neighboring spheres almost touch). With this $n = 1/a^3 = 1/8R^3$, Eq. (33) yields $\varepsilon_r = 1 + 4\pi/8 \approx 2.5$. Due to the crude nature of this estimate, we may conclude that it provides a reasonable explanation for the values of ε_r , listed in first few lines of Table 1. Still, it is clear that such model cannot even approximately describe dielectric properties of either water or barium titanate (and similar materials), as well as their strong temperature dependence. Such high values may be explained by the *molecular field effect*: each elementary dipole is polarized not only by the external field (as in our current toy model), but by the field of neighboring dipoles as well.

Before analyzing this effect (in the next section), let us review how are the most important results of electrostatics modified by a *uniform* linear dielectric medium that obeys Eq. (38) with a space-independent dielectric constant ε_r . The simplest problem of this kind is a set of free charges of density $\rho(\mathbf{r})$, inserted into the medium. For this case, we can combine Eqs. (32) and (38) to write

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon}, \quad \text{i.e. } \nabla^2 \phi = -\frac{\rho}{\varepsilon}. \quad (3.43)$$

For charges in vacuum, we had similar equations (1.27) and (1.41), but with a different constant, $\varepsilon_0 = \varepsilon/\varepsilon_r$. Hence all the results discussed in Chapter 1 are valid, with both \mathbf{E} and ϕ reduced by the factor of ε_r . Thus, the most straightforward result of the induced polarization of a dielectric media is the electric field reduction. This is a very important effect, especially taken into account the very high values of ε_r in such dielectrics as water – see Table 1. Indeed, this is the reduction of the attraction between positive

¹⁵ Note that for all materials listed in Table 1, $\varepsilon_r > 1$, i.e. $\alpha_{\text{mol}} > 0$. Actually, this is true for all stable dielectrics. Let me postpone a discussion of this fact until Sec. 5.5 where I will compare physical mechanisms of the electric and magnetic polarization.

¹⁶ A more accurate model of atomic polarization is discussed in QM Chapter 6.

¹⁷ See, e.g., SM Secs. 1.4 and 3.1.

¹⁸ Such radius is defined by the requirement that the volume of the corresponding sphere, used in the van-der-Waals equation (see, e. g., SM Sec. 4.1) gives the best fit to the experimental equation of state $n = n(P, T)$.

and negative ions (called, respectively, *cations* and *anions*) in water that enables their substantial dissociation and hence almost all biochemical reactions, which are the basis of biological cell functions - and hence of the life itself.

Now, what if the electric field in a uniform dielectric is induced by charges located on conductors - with potentials rather than charge density fixed? Then, with the substitution of the electrostatic potential definition $\mathbf{E} \equiv -\nabla\phi$, Eq. (43) in the space between the conductors is reduced to the Laplace equation, and the boundary problem remains exactly the same as formulated in Chapter 2 – see Eqs. (2.35). Hence the potential distribution $\phi(\mathbf{r})$ is related to the conductor potential in exactly the same way as in vacuum (see, e.g., any problem discussed in Chapter 2), *without any effect* of the medium polarization. However, in order to find, from that distribution, the density σ of charges on conductor surfaces, we need to use the macroscopic Gauss law (34). Applying this equation to a pillbox-shaped volume on the conductor surface, we get the following relation,

$$\sigma = D_n = \varepsilon E_n = -\varepsilon \frac{\partial \phi}{\partial n}, \quad (3.44)$$

which differs from Eq. (2.3) only by the replacement $\varepsilon_0 \rightarrow \varepsilon = \varepsilon_r \varepsilon_0$. Hence the charge density, calculated for the vacuum case, should be increased by the factor of ε_r – that's it. In particular, this means that all the capacitances that had been calculated in vacuum, should be increased by that factor. For example, for planar capacitor filled with linear dielectric ε_r , we get the well-known formula

$$C_m = \frac{\varepsilon_r \varepsilon_0 A}{d} = \frac{\varepsilon A}{d}. \quad (3.45)$$

C_m of a
planar
capacitor

(As a reminder, this increase of C_m by ε_r has been already used in Sec. 2.2 for capacitance estimates.)

Now let us discuss more complex situations in which the dielectric medium is *not* uniform, for example when it contains a boundary separating two regions filled by different uniform dielectrics. (The analysis is clearly applicable to a dielectric/vacuum boundary as well, with one of the dielectric constants set to 1.) For that, let us apply the macroscopic Gauss law (34) to a pillbox formed at the interface between two dielectrics, with no surface charges – see the solid lines in Fig. 5.

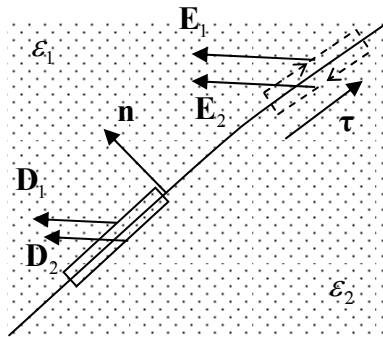


Fig. 3.5. Deriving boundary conditions on the interface between two dielectrics: a Gauss pillbox and a circulation contour. \mathbf{n} and $\boldsymbol{\tau}$ are the unit vectors which are, respectively, normal and tangential to the interface.

This immediately gives $(D_n)_1 = (D_n)_2$, so that Eq. (38) yields

$$(\varepsilon E_n)_1 = (\varepsilon E_n)_2, \quad \text{i.e.} \quad \varepsilon_1 \frac{\partial \phi_1}{\partial n} = \varepsilon_2 \frac{\partial \phi_2}{\partial n}. \quad (3.46)$$

Boundary
condition
for E_n

Now, what about the tangential component (E_τ) of the electric field? In dielectrics, static electric field is still potential, hence we can still use Eq. (1.28). Integrating this equation along to a narrow contour stretched along the interface (see the dashed line in Fig. 5), we get

$$(E_\tau)_1 = (E_\tau)_2, \quad \text{i.e.} \quad \frac{\partial \phi_1}{\partial \tau} = \frac{\partial \phi_2}{\partial \tau}. \quad (3.47)$$

Boundary
condition
for E_τ

Note that this condition is compatible with (and may be derived follows from) the continuity of the electrostatic potential itself, $\phi_1 = \phi_2$, at each point of the interface. That relation may be derived from the electric field definition as the gradient of ϕ - see Eq. (1.33). Indeed, if the potential leaped at the border, the electric field would be infinite.

Let us apply the boundary conditions (46)-(47), for example, to two thin ($t \ll d$) vacuum slits cut in a uniform dielectric with an initially uniform¹⁹ electric field \mathbf{E}_0 (Fig. 6). In both cases, a slit with $t \rightarrow 0$ cannot modify the field distribution outside it substantially.

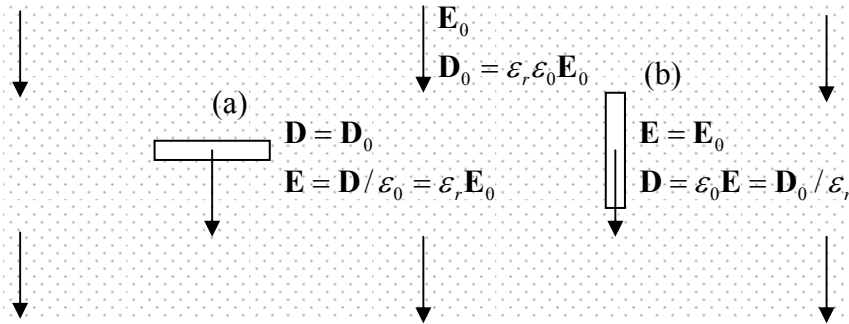


Fig. 3.6. Fields inside narrow slits cut in a linear dielectric.

For slit (a), normal to the applied field, we may apply Eq. (46) to the “major” (broad) interfaces, shown horizontal in Fig. 6, we see that D should be continuous. But according to Eq. (33), this means that inside the gap (i. e. in the vacuum, with $\mathbf{P} = 0$) the electric field equals \mathbf{D}/ϵ_0 . This field, and hence \mathbf{D} , may be measured, showing that the electric displacement is not a purely mathematical construct. Superficially, this result violates the boundary condition (47) on the vertical (“minor”) interfaces of the slit. Note, however, that the electric field within the gap is ϵ_r times higher than in the dielectric outside it. Hence the slit deforms the equipotential surfaces around it to concentrate the field inside itself. The curving of the surfaces near the minor interfaces takes care of the fulfillment of Eq. (47) at the minor interfaces.

On the contrary, for slit (b) parallel to the applied field, we may apply Eq. (47) to the major (now, vertical) interfaces of the slit, to see that it is electric field \mathbf{E} that is continuous now, while the electric displacement $\mathbf{D} = \epsilon_0 \mathbf{E}$ inside the gap is a factor of ϵ_r lower than its value in the dielectric. (Any perturbation of the field uniformity, caused by the compliance with Eq. (46) at the minor interfaces, is settled at distances $\sim t$ from these interfaces.)

¹⁹ Actually, selecting the slit size d much less than the characteristic scale of the field change, we can apply the following arguments to *any* external field distribution.

For problems with piecewise-constant ε but more complex geometries we may need to apply the methods studied in Chapter 2. As in vacuum, in the simplest cases we can select such a set of orthogonal coordinates that the electrostatic potential depends on just one of them. Consider, for example, two types of plane capacitor filling with two different dielectrics – see Fig. 7.

In case (a), voltage V between the electrodes is the same for each part of the capacitor, and at least far from the dielectric interface, the electric field is vertical, uniform, and similar ($E = V/d$). Hence the boundary condition (47) is satisfied even if such a distribution is valid near the surface as well, i.e. at any point of the system. The only effect of different values of ε in the two parts is that the electric displacement $D = \varepsilon E$ and hence electrodes' surface charge density $\sigma = D$ are different in the two parts. Thus we can calculate the electrode charges $Q_{1,2}$ of the two parts independently, in each case using Eq. (44), and then add up the results to get the total capacitance

$$C_m = \frac{Q_1 + Q_2}{V} = \frac{1}{d}(\varepsilon_1 A_1 + \varepsilon_2 A_2). \quad (3.48)$$

Note that this formula may be interpreted as the total capacitance of two separate capacitors connected (by conducting wires) *in parallel*. This is natural, because we may cut the system along the dielectric interface, without any effect on the fields in either part, and then connect the corresponding electrodes by external wires, again without any effect on the system – besides very close to capacitor's edges.

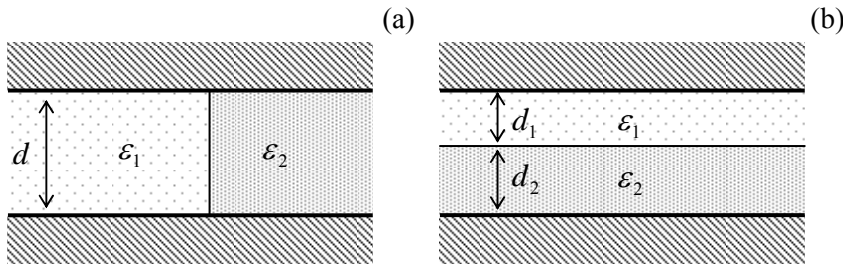


Fig. 3.7. Plane capacitors filled with two different dielectrics.

Case (b) may be analyzed by applying Eq. (34) to a Gaussian pillbox with the lower lid inside the (for example) bottom electrode, and the top lid in any of the layers. From this we see that D anywhere inside the system should be equal to the surface charge density σ of the lower electrode, i.e. constant. Hence, in the top dielectric layer the electric field is constant: $E_1 = D_1/\varepsilon_1 = \sigma/\varepsilon_1$, while in bottom layer, similarly, $E_2 = D_2/\varepsilon_2 = \sigma/\varepsilon_2$. Integrating E across the whole capacitor, we get

$$V = \int_0^{d_1+d_2} E(z) dz = E_1 d_1 + E_2 d_2 = \left(\frac{d_1}{\varepsilon_1} + \frac{d_2}{\varepsilon_2} \right) \sigma, \quad (3.49)$$

so that the mutual conductance per unit area

$$\frac{C_m}{A} \equiv \frac{\sigma}{V} = \left[\frac{d_1}{\varepsilon_1} + \frac{d_2}{\varepsilon_2} \right]^{-1}. \quad (3.50)$$

Note that this result is equivalent to the total capacitance of a *series* connection of two plane capacitors based on each of the layers. This is natural, because we could insert an uncharged thin conducting sheet (rather than a cut as in the previous case) at the layer interface, which is an

equipotential surface, without changing the field distribution in the system. Then we could thicken the conducting sheet as much as we like (turning it into a wire), also without changing the fields and hence the capacitance.

In order to warm up for more complex problems, let us see how the last problem could be solved using the Laplace equation approach. Due to the symmetry of the system, the electrostatic potential in each layer may only depend on one (in Fig. 7b, vertical) coordinate z , so that the Laplace equation in each uniform part of the system is reduced to $d^2\phi/d^2z = 0$. Hence in each layer the electrostatic potential changes linearly, though possibly with different coefficients: $\phi_1 = c_{11}z + c_{12}$, and $\phi_2 = c_{21}z + c_{22}$. Selecting the electrode potentials as $\phi(0) = 0$ and $\phi(d_1 + d_2) = V$, from those boundary conditions we get $c_{12} = 0$, $c_{21}(d_1 + d_2) + c_{22} = V$, so that we need two more equations to find all four coefficients c_{ij} . These additional equations come from the conditions of continuity of the potential ($c_{11}d_1 + c_{12} = c_{21}d_1 + c_{22}$) and displacement ($\epsilon_1 c_{11} = \epsilon_2 c_{21}$) at the interface $z = d_1$. Solving these equations, we can readily find the electric field and displacement in both layers, then the surface charge densities

$$\sigma(0) = D|_{z=0} = -\epsilon_1 \frac{d\phi_1}{dz}|_{z=0}, \quad \sigma(d_1 + d_2) = D|_{z=d_1+d_2} = -\epsilon_2 \frac{d\phi_2}{dz}|_{z=d_1+d_2} \quad (3.51)$$

(which in this case are equal and opposite) and finally the capacitance per unit area, with (of course) the same result (50).

Let us apply the same approach to a more complex problem, shown in Fig. 8a: a dielectric sphere placed into a uniform external electric field \mathbf{E}_0 .

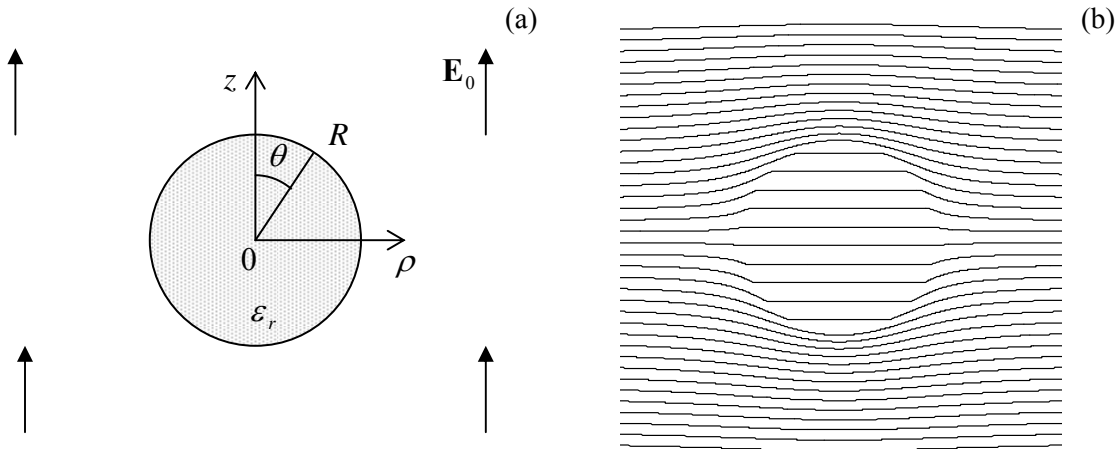


Fig. 3.8. Dielectric sphere in an initially uniform electric field: (a) the problem, and (b) the equipotential surfaces, as given by Eq. (55), for $\epsilon_r = 3$.

In this case the Laplace equation is not one-dimensional, and hence invites the variable separation method discussed in Sec. 2.5. From that discussion we already know, in particular, the general solution (2.172) of the Laplace equation outside of the sphere. To satisfy the uniform-field condition at $r \rightarrow \infty$, it reduces to

$$\phi_{r \geq R} = -E_0 r \cos \theta + \sum_{l=1}^{\infty} \frac{b_l}{r^{l+1}} \mathcal{P}_l(\cos \theta). \quad (3.52)$$

Inside the sphere we can only use the radial functions that are finite at $r \rightarrow 0$:

$$\phi_{r \leq R} = \sum_{l=1}^{\infty} a_l r^l \mathcal{P}_l(\cos \theta). \quad (3.53)$$

Now, writing the boundary conditions (46) and (47) at $r = R$, we see that for all coefficients a_l and b_l with $l \geq 2$ we (just like for the conducting sphere in vacuum) get homogeneous equations that have only trivial solutions. Hence, all these terms may be dropped, while for the only surviving angular harmonic, proportional to $\mathcal{P}_1(\cos \theta) = \cos \theta$, Eqs. (46)-(47) yield two equations:

$$-E_0 - \frac{2b_1}{R^3} = \varepsilon_r a_1, \quad -E_0 R + \frac{b_1}{R^2} = a_1 R. \quad (3.54)$$

Solving this simple system for a_1 and b_1 , we get the final solution of the problem:

$$\phi_{r \geq R} = E_0 \left(-r + \frac{\varepsilon_r - 1}{\varepsilon_r + 2} \frac{R^3}{r^2} \right) \cos \theta, \quad \phi_{r \leq R} = -E_0 \frac{3}{\varepsilon_r + 2} r \cos \theta. \quad (3.55)$$

Figure 8b shows the equipotential surfaces given by this solution, for a particular value off the dielectric constant ε_r . Note that, just like for a conducting sphere, at $r \geq R$ the dielectric sphere produces (on the top of the uniform external field) a purely dipole field, with $\mathbf{p} = 4\pi R^3 \varepsilon_0 \mathbf{E}_0 (\varepsilon_r - 1)/(\varepsilon_r + 2)$ – an evident generalization of Eq. (11), to which our result tends at $\varepsilon_r \rightarrow \infty$. By the way, this property is common: from the point of view of their electrostatic (but not transport!) properties, conductors may be adequately described as dielectrics with $\varepsilon_r \rightarrow \infty$.

Another remarkable feature of Eqs. (55) is that the electric field inside the sphere is uniform²⁰ with R -independent values

$$\mathbf{E} = \frac{3}{\varepsilon_r + 2} \mathbf{E}_0, \quad \mathbf{D} \equiv \varepsilon_r \varepsilon_0 \mathbf{E} = \varepsilon_0 \frac{3\varepsilon_r}{\varepsilon_r + 2} \mathbf{E}_0, \quad \mathbf{P} \equiv \mathbf{D} - \varepsilon_0 \mathbf{E} = 3\varepsilon_0 \frac{\varepsilon_r - 1}{\varepsilon_r + 2} \mathbf{E}_0. \quad (3.56)$$

In the limit $\varepsilon_r \rightarrow 1$ (the “vacuum sphere”, i.e. no sphere at all), the electric field inside the sphere naturally tends to the external one, and its polarization disappears. In the opposite limit and $\varepsilon_r \rightarrow \infty$ the electric field inside the sphere vanishes, and the field outside the sphere approaches that we have found for the conducting sphere – see Eq. (2.176).

To complete the discussion of this example, note a very curious result: the field \mathbf{E}_{self} , created by the dielectric sphere inside itself, is related to its polarization vector by a simple equation independent of either the dielectric constant or sphere’s size:

$$\mathbf{E}_{\text{self}} \equiv \mathbf{E} - \mathbf{E}_0 = -\frac{\varepsilon_r - 1}{\varepsilon_r + 2} \mathbf{E}_0 = -\frac{1}{3\varepsilon_0} \mathbf{P}, \quad (3.57)$$

where factor 3 stems sphere’s dimensionality. (For a round cylinder in a normal external field, the similar relation is valid, but with factor 2.) This equality is just the particular manifestation of the general relation (24). Indeed, if summed over all $N = nV$ similar dipoles \mathbf{p} , distributed inside the sphere with constant density \mathbf{n} (so that the polarization vector $\mathbf{P} = n\mathbf{p}$ is constant), Eq. (24) yields

²⁰ This is true for any ellipsoid, at arbitrary external field orientation.

$$\int_V \mathbf{E}_{\text{self}}(\mathbf{r}) d^3r = -\frac{\mathbf{P}}{3\epsilon_0} V, \quad (3.58)$$

so that after division by V , and taking into account the field uniformity in our particular case, it coincides with Eq. (57).²¹ We will use these results in the following section to discuss the molecular field effect.

Before doing that, let me briefly revisit the method of charge images that was discussed in Sec. 2.6, to find its new features pertaining to dielectrics. As the simplest example, consider a point charge near a dielectric half-space – see Fig. 9 (cf. Fig. 2.24).

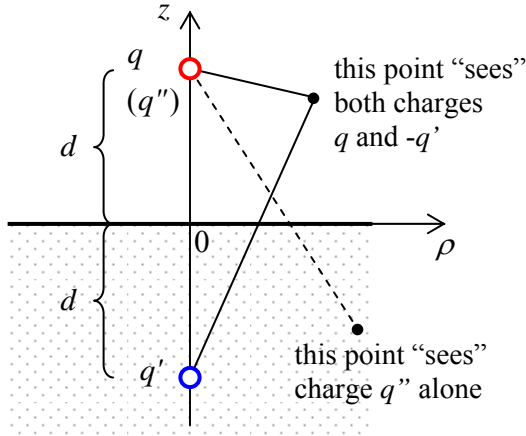


Fig. 3.9. Charge images for a dielectric half-space.

The Laplace equations in the upper half-space $z > 0$ (besides the charge point $\rho = 0, z = d$) may still be satisfied using a single image charge q' at point $\rho = 0, z = -d$, but now q' may differ from $(-q)$. In addition, in contrast to the conducting plane case, we should also find the field inside the dielectric ($z \leq 0$). This field cannot be contributed by the image charge, because it would provide a potential divergence at its location. Thus, in that half-space we should try to use the real point source only, but maybe with a re-normalized charge q'' rather than the genuine charge q – see Fig. 9. As a result, we may look for the potential distribution in the form

$$\phi(\rho, z) = \frac{1}{4\pi\epsilon_0} \times \begin{cases} \left[\frac{q}{(\rho^2 + (z-d)^2)^{1/2}} + \frac{q'}{(\rho^2 + (z+d)^2)^{1/2}} \right], & \text{for } z \geq 0, \\ \frac{q''}{(\rho^2 + (z-d)^2)^{1/2}}, & \text{for } z \leq 0, \end{cases} \quad (3.59)$$

at this stage with unknown q' and q'' . Plugging this solution into the boundary conditions (46) and (47) at $z = 0$ (with $\partial/\partial n = \partial/\partial z$), we see that they are indeed satisfied (so that Eqs. (59) express the unique solution of the boundary problem) if the effective charges q' and q'' obey the following relations:

²¹ The reader may wonder how have we managed to proof Eq. (24), at least for this particular case, using only the relations based on the dipole approximation (7) for the field, which does not cover the inter-dipole fields responsible for Eq. (24) – see Fig. 3 and its discussion. The reason is that according to Eq. (30), the additional field \mathbf{E}_{self} inside the sphere may be considered as been created by effective charges, of density ρ_{ef} , distributed on sphere's surface. For these charges, field \mathbf{E}_{ef} is *internal*, similar to the field between two charges, shown in Fig. 3.

$$q - q' = \varepsilon_r q'', \quad q + q' = q''. \quad (3.60)$$

Solving this simple system of linear equations, we get

$$q' = -\frac{\varepsilon_r - 1}{\varepsilon_r + 1} q, \quad q'' = \frac{2}{\varepsilon_r + 1} q. \quad (3.61)$$

If $\varepsilon_r \rightarrow 1$, then $q' \rightarrow 0$, and $q'' \rightarrow q$ – both facts very natural, because in this limit (no polarization!) we have to recover the unperturbed field of the initial point charge in both semi-spaces. In the opposite limit $\varepsilon_r \rightarrow \infty$ (which, according to our discussion of the last problem, should correspond to a conducting plane), $q' \rightarrow q$ (repeating the result we have discussed in very much detail in Sec. 2.6) and $q'' \rightarrow 0$. According to the second of Eqs. (3.59), the last result means the field in the dielectric tends to zero in this limit, as it should.

Finally, following the logic of Chapter 2, at this point it would be appropriate to discuss the Green's function method. However, due to the time/space restrictions, I will skip this discussion, especially because the all the method's philosophy remains absolutely the same as for the vacuum case, so that the generalization to the case of dielectrics is straightforward.

3.4. Molecular field effects

In 1850, O.-F. Mossotti and (probably, independently, but almost 30 years later!) R. Clausius made an interesting experimental observation known now, rather unfairly, as the *Clausius-Mossotti relation*: if density n of molecules in a chemical compound may be changed without changing its molecular structure, then the following ratio,

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2}, \quad (3.62)$$

is approximately proportional to n . For $\varepsilon_r \rightarrow 1$, i.e., $n \rightarrow 0$, there is no surprise here: according to Eq. (41), for independent molecular dipoles $\varepsilon_r - 1 = 4\pi\alpha_{mol}n \propto n$. However, at larger density n , the effective field \mathbf{E}_{ef} , acting on each dipole, includes not only the external field \mathbf{E}_0 , but also a substantial “molecular field” \mathbf{E}_{mol} of the surrounding dipoles:

$$\mathbf{E}_{ef} = \mathbf{E}_0 + \mathbf{E}_{mol}(0), \quad (3.63)$$

where the position of the particular dipole we are discussing is taken for $\mathbf{r} = 0$. Let us calculate $\mathbf{E}_{mol}(0)$, using a very simple model: a regular cubic lattice of identical dipoles (Fig. 10). In a Cartesian coordinate system with axes directed along the lattice vectors, coordinates of the dipoles are

$$x_{jkl} = aj, \quad y_{jkl} = ak, \quad z_{jkl} = al, \quad (3.64)$$

where j , k , and l are the integers numbering the dipoles. Now we may use the last form of Eq. (13), and the linear superposition principle, to calculate one of the Cartesian components (say, along axis x) of the molecular field induced by all other dipoles of the lattice:

$$(E_{mol})_x(0) = \frac{1}{4\pi\varepsilon_0 a^3} \sum_{j,k,l=-\infty}^{+\infty} \frac{3j(jp_x + kp_y + lp_z) - p_x(j^2 + k^2 + l^2)}{(j^2 + k^2 + l^2)^{5/2}}, \quad (3.65)$$

with excluded term $j = k = l = 0$ is excluded. The sums of all cross-terms, proportional to jk and jl , vanish due to system symmetry, so that Eq. (65) reduces to

$$(E_{\text{mol}})_x(0) = \frac{1}{4\pi\epsilon_0 a^3} \sum_{j,k,l=-\infty}^{+\infty} \frac{[3j^2 - (j^2 + k^2 + l^2)]}{(j^2 + k^2 + l^2)^{5/2}} p_x. \quad (3.66)$$

Since all the sums participating in this expression are equal,

$$\sum_{j,k,l=-\infty}^{+\infty} \frac{j^2}{(j^2 + k^2 + l^2)^{5/2}} = \sum_{j,k,l=-\infty}^{+\infty} \frac{k^2}{(j^2 + k^2 + l^2)^{5/2}} = \sum_{j,k,l=-\infty}^{+\infty} \frac{l^2}{(j^2 + k^2 + l^2)^{5/2}}, \quad (3.67)$$

we get $(E_{\text{mol}})_x(0) = 0$. Due to the system symmetry, the same result is valid for all other components of the dipole field. Hence, $\mathbf{E}_{\text{mol}}(0) = 0$, and (due to the equivalence of all the dipoles of the system), the molecular field vanishes at the location of each dipole, so that Eq. (3.63) is reduced to $\mathbf{E}_{\text{ef}} = \mathbf{E}_0$.

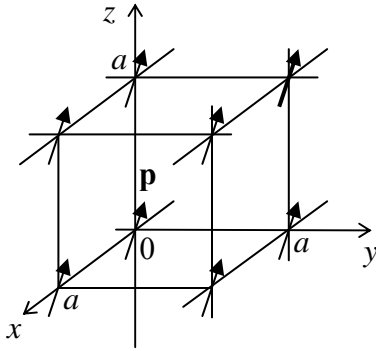


Fig. 3.10. Cubic lattice of similar dipoles.

In order to relate the external field \mathbf{E}_0 and the average dipole²² field \mathbf{E} in the medium, we may use Eq. (56) for a uniform, macroscopic sphere²³ with a radius much larger than the inter-dipole distance a , so that our assumption of infinite limits of the rapidly converging sum (65) is not substantially affected:

$$\mathbf{E} = \frac{3}{\epsilon_r + 2} \mathbf{E}_0 = \frac{3}{\epsilon_r + 2} \mathbf{E}_{\text{ef}}. \quad (3.68)$$

Now we may plug this relation into the general formula (37) for linear dielectrics:

$$\mathbf{P} = (\epsilon_r - 1)\epsilon_0 \mathbf{E} = \frac{3(\epsilon_r - 1)}{\epsilon_r + 2} \epsilon_0 \mathbf{E}_{\text{ef}}. \quad (3.69)$$

This “macroscopic” relation has to give the same result as the “microscopic” Eq. (40) - with the replacement $\mathbf{E} \rightarrow \mathbf{E}_{\text{ef}}$ which reflects the fact that in the general case each dipole is polarized by the effective field (63) rather than the average field \mathbf{E} :

$$\mathbf{P} = 4\pi\epsilon_0 \alpha_{\text{mol}} n \mathbf{E}_{\text{ef}}. \quad (3.70)$$

²² This qualifier is important: \mathbf{E} is the long-range (dipole field) average participating in the macroscopic Maxwell equations, rather than the exact average that would include the inner-dipole fields, for which Eq. (24) would be valid.

²³ This geometry, due to its isotropy, most fairly represents the relation between \mathbf{E} and \mathbf{E}_0 .

Lorentz-
Lorenz
formula

The comparison yields the so-called *Lorentz-Lorenz* formula,²⁴

$$\frac{4\pi\alpha_{\text{mol}}}{3}n = \frac{\varepsilon_r - 1}{\varepsilon_r + 2}, \quad (3.71)$$

that complies with the Clausius-Mossotti observation, provided that the molecular polarizability α_{mol} is independent of density. (This is a good approximation at least for weak “molecular” bonding.)

It is somewhat surprising how many dielectric materials obey Eq. (71) rather well, because of its approximate nature. Indeed, its derivation is based on the assumption of a specific crystal lattice and, more importantly, that the molecules are localized exactly in the crystal lattice nodes, and the field of each molecule may be expressed by the dipole approximation. In reality, atom’s electrons, which participate in the dipole moment formation, are spread in space due to quantum-mechanical uncertainty on a scale that may be comparable with distances between the molecules.

Solving Eq. (71) for the dielectric constant, we get

$$\varepsilon_r = \frac{1 + 8\pi\alpha_{\text{mol}}n/3}{1 - 4\pi\alpha_{\text{mol}}n/3}. \quad (3.72)$$

If the dipole density is low, $\alpha_{\text{mol}}n \ll 1$, we get our old result (41) corresponding to independent dipoles, and hence to $\mathbf{E}_{\text{ef}} = \mathbf{E}$. However, at high dipole density and/or polarizability, the effective field acting on the each dipole,

$$\mathbf{E}_{\text{ef}} = \frac{\varepsilon_r + 2}{3}\mathbf{E} = \frac{\mathbf{E}}{1 - 4\pi\alpha_{\text{mol}}n/3}, \quad (3.73)$$

may be substantially larger than the average field \mathbf{E} , due to the molecular field contribution. Note ε_r , the $\mathbf{E}_{\text{ef}}/\mathbf{E}$ ratio, and hence the electric susceptibility

$$\chi_e \equiv \frac{\mathbf{P}}{\varepsilon_0\mathbf{E}} = \varepsilon_r - 1 = \frac{4\pi\alpha_{\text{mol}}n}{1 - 4\pi\alpha_{\text{mol}}n/3}, \quad (3.74)$$

all diverge when the density-polarizability product approaches the critical value $\alpha_{\text{mol}}n = 3/4\pi$.

This is essentially a rudimentary²⁵ description of the transition from linear dielectrics to the so-called *ferroelectrics* with self-sustained (*spontaneous*) polarization even in the absence of external

²⁴ It was derived by in 1869 by L. Lorenz and then (in 1878) independently by H. Lorentz. Actually, they discussed optical frequencies at which ε_r should be understood as the square of the refraction coefficient at the wave frequency (see Chapter 7), but since the optical wavelengths $\sim 10^{-4}$ m are much longer than interatomic distances $a \sim 10^{-9}$ m, the derivation remains absolutely the same in electrostatics.

²⁵ Any quantitative description of this transition should account of for thermal fluctuations of the molecular dipoles, which reduce the dipole-dipole ordering and hence suppress the transition to the ferroelectric phase until temperature has been lowered to a certain *Curie temperature* T_C - named after P. Curie (1859-1906). Right above that temperature, the dielectric remains linear, but has a high, temperature-dependent dielectric constant that diverges at $T \rightarrow T_C$. Such materials are frequently called *paraelectric*, and the paraelectric-to-ferroelectric transition at T_C in crystals is a typical example of a *continuous* (or “second-order”) *phase transition* - see, e.g., SM Sec. 4.4. (As will be discussed in Sec. 5.5 below, some magnetic materials exhibit a very similar phase transition between their *ferromagnetic* and *paramagnetic* phases.) Moreover, in non-crystalline materials, such as bulk ceramics and thin films, the ferroelectric behavior is further complicated by different, field-dependent

electric field. These materials are typically recognized by the hysteretic behavior of their polarization as a function of applied electric field – see, for example, Fig. 11.

Ferroelectric materials are being actively explored as the active materials for nonvolatile random-access memories (dubbed either FRAM or FeRAM).²⁶ In cells of this memory, binary information is stored in the form of one of two possible directions of spontaneous polarization at $\mathbf{E} = 0$ – see, e.g., Fig. 11, and is read out by the effect of the average electric field on a nearby semiconductor field-effect transistor. Unfortunately, most materials suitable for fabrication of ferroelectric thin films are rather complex and incompatible with standard processes of microelectronics. In addition, the time of spontaneous depolarization of ferroelectric thin films is typically well below than 10 years – the industrial standard for data retention in nonvolatile memories, and this time may be decreased even more by “fatigue” from repeated polarization recycling. Due to these reasons, industrial production of FRAM is currently just a tiny, few-percent fraction of the nonvolatile memory market (which is currently dominated by floating-gate memories – see Sec. 4.2).

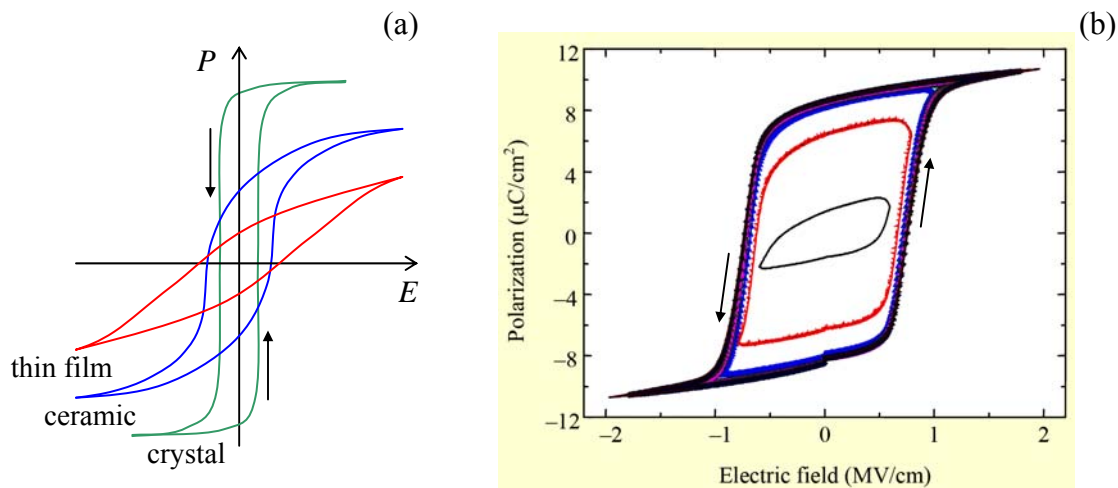


Fig. 3.11. Ferroelectric hysteric loops: (a) for various material types (schematically), and (b) for several amplitudes of the applied ac electric field. (Panel b, showing recent (2013) experimental results by S.-W. Jung *et al.* for an inkjet-printed layer of organic semiconductor PC12TV12T, is adapted from <http://etrij.etri.re.kr/etrij/journal/article/article.do?volume=35&issue=4&page=734>.)

Other polarization effects can also be met, possible, e.g., *antiferroelectricity* or *helielectricity*. Unfortunately, we will not have time for a discussion of these exotic phenomena in this course;²⁷ the main reason I am mentioning them is to emphasize again that the “material relation” $\mathbf{P} = \mathbf{P}(\mathbf{E})$ is by no means exact or fundamental, though most material, in practicable fields, behave as linear dielectrics.

directions of polarization \mathbf{P} in individual “domains” of the sample, making the average hysteresis more smooth (Fig. 11a) and dependent on sample’s polarization history – for example the amplitude of the applied ac electric field (Fig. 11b).

²⁶ See, e.g., J. F. Scott, *Ferroelectric Memories*, Springer, 2000.

²⁷ For a detailed coverage of ferroelectrics, I can recommend an encyclopedic monograph by M. Lines and A. Glass, *Principles and Applications of Ferroelectrics and Related Materials*, Oxford U. Press, 2001, and a recent collection of reviews by K. Rabe, C. Ahn, and J.-M. Triscone (eds.), *Physics of Ferroelectrics: A Modern Perspective*, Springer, 2010.

3.5. Energy of electric field in a dielectric

In Chapter 1, we have obtained two key results for the electrostatic energy: Eq. (1.54) for a charge interaction with an independent (“external”) field, and a similarly structured formula (1.62), but with an additional factor $\frac{1}{2}$, for the field is produced by the charges under consideration. Both relations could be merged and rewritten in a “local” form involving energy density u – see Eq. (1.67). These equations are of course always valid for dielectrics as well if the charge density includes *all* charges (including those bound into dipoles), but it is convenient to recast them unto a form depending on density $\rho(\mathbf{r})$ of only “stand-alone” charges.

If a field is created only by stand-alone charges under consideration, and is proportional to $\rho(\mathbf{r})$ (requiring that we deal with a *linear* dielectric!), we can repeat all the argumentation of the beginning of Sec. 1.3, and again arrive at Eq. (1.62), provided that ϕ is calculated correctly, i.e., with a due account of the dielectric. Now we can recast this result in terms of fields – essentially as this was done in Eqs. (1.64)-(1.66), but now making a clear difference between the electric field \mathbf{E} (that still equals $-\nabla\phi$) and the electric displacement field \mathbf{D} that obeys the macroscopic Maxwell equation (32). Plugging $\rho(\mathbf{r})$, expressed from that equation, into Eq. (1.62), we get

$$U = \frac{1}{2} \int (\nabla \cdot \mathbf{D}) \phi \, d^3 r. \quad (3.75)$$

Using the fact²⁸ that for any differentiable functions ϕ and \mathbf{D} ,

$$(\nabla \cdot \mathbf{D}) \phi = \nabla \cdot (\phi \mathbf{D}) - (\nabla \phi) \cdot \mathbf{D}, \quad (3.76)$$

we may rewrite Eq. (75) as

$$U = \frac{1}{2} \int \nabla \cdot (\phi \mathbf{D}) \, d^3 r - \frac{1}{2} \int (\nabla \phi) \cdot \mathbf{D} \, d^3 r. \quad (3.77)$$

The divergence theorem, applied to first term, reduces it to a surface integral of ϕD_n . (As a reminder, in Eq. (1.65) the integral was of $\phi(\nabla \phi)_n \propto \phi E_n$.) If the surface of the volume we consider is sufficiently far, this surface integral vanishes. On the other hand, the gradient in the second term of Eq. (77) is just (minus) field \mathbf{E} , so that it gives

$$U = \frac{1}{2} \int \mathbf{E} \cdot \mathbf{D} \, d^3 r = \frac{1}{2} \int E(\mathbf{r}) \cdot \varepsilon(\mathbf{r}) E(\mathbf{r}) \, d^3 r = \frac{\varepsilon_0}{2} \int \varepsilon_r(\mathbf{r}) E^2(\mathbf{r}) \, d^3 r. \quad (3.78)$$

This expression is a natural generalization of Eq. (1.67) and shows that we can, like we did in vacuum, present the electrostatic energy in a local form²⁹

Field
energy in
a linear
dielectric

$$U = \int u(\mathbf{r}) \, d^3 r, \quad u = \frac{1}{2} \mathbf{E} \cdot \mathbf{D} = \frac{\varepsilon}{2} E^2 = \frac{D^2}{2\varepsilon}. \quad (3.79)$$

Again, this expression is not valid for *nonlinear* dielectrics, because our starting point, Eq. (1.62), is only valid if ϕ is proportional to ρ . In order to make our calculation more general, we should

²⁸ See, e.g., MA Eq. (11.4a).

²⁹ Again, in Gaussian units this expression should be divided by 4π .

intercept our calculations in Sec. 1.3 at an earlier stage, at which we have not yet used this proportionality. For example, Eq. (1.54) may be rewritten, in the continuous limit, as

$$\delta U = \int \phi(\mathbf{r}) \delta \rho(\mathbf{r}) d^3 r, \quad (3.80)$$

where symbol δ means a small variation of the function - e.g., its change in time, sufficiently slow to ignore the relativistic and magnetic-field effects. Applying such variation to Eq. (32), and plugging the resulting $\delta \rho = \nabla \cdot \delta \mathbf{D}$ into Eq. (80), we get

$$\delta U = \int (\nabla \cdot \delta \mathbf{D}) \phi d^3 r. \quad (3.81)$$

(Note that in contrast to Eq. (75), this expression does not have factor $\frac{1}{2}$.) Now repeating the same calculations as in the linear case, for the energy density *variation* we get a remarkably simple (and general!) expression,

$$\delta u = \mathbf{E} \cdot \delta \mathbf{D}. \quad (3.82)$$

General
energy
variation

This is as far as we can go for the general dependence $\mathbf{D}(\mathbf{E})$. If the dependence is linear and isotropic, as in Eq. (38), then $\delta \mathbf{D} = \epsilon \delta \mathbf{E}$ and

$$\delta u = \epsilon \mathbf{E} \cdot \delta \mathbf{E} \quad (3.83)$$

Integration of this expression over variations, from zero field to a certain final distribution $\mathbf{E}(\mathbf{r})$, brings us back to Eq. (79).

Another important role of Eq. (82) is that it shows that Cartesian coordinates of \mathbf{E} may be interpreted as generalized forces, and those of \mathbf{D} as generalized coordinates (per unit volume).³⁰ This allows one to form the proper *Gibbs potential energy*³¹ of a system inside some volume V , placed in an external electric field \mathbf{E}_{ext} :

$$\mathcal{G} = \int_V g(\mathbf{r}) d^3 r, \quad g(\mathbf{r}) = u(\mathbf{r}) - \mathbf{E}_{\text{ext}} \cdot \mathbf{D}. \quad (3.84)$$

Gibbs
potential
energy

As an analytical mechanics reminder, if a generalized external force (in our case, \mathbf{E}_{ext}) is fixed, the stable equilibrium of the system corresponds to the minimum of \mathcal{G} , rather than of the potential energy U as such - in our case, that of the field:

$$U = \int_V u(\mathbf{r}) d^3 r, \quad u(\mathbf{r}) = \int \mathbf{E} \cdot \delta \mathbf{D}. \quad (3.85)$$

In order to illustrate this important point, let us return to the simple case of a system with linear dielectric(s), in which $\delta \mathbf{D} \propto \delta \mathbf{E} \propto \delta \mathbf{E}_{\text{ext}}$, so that Eq. (85) may be explicitly integrated over the external field variation, to reproduce the second of Eqs. (79):

³⁰ This is the point where the SI units, prescribing fields \mathbf{E} and \mathbf{D} different dimensionalities, are more revealing than the Gaussian units.

³¹ See, e.g., CM Sec. 1.5. Note that as Eq. (84) clearly illustrates, once again, that the difference between potential energies \mathcal{G} and U , usually discussed in courses of statistical physics and/or thermodynamics as the difference between the Gibbs and Helmholtz free energies (see, e.g., SM 1.6), exists regardless of statistics or thermal motion.

$$u(\mathbf{r}) = \frac{1}{2} \mathbf{E} \cdot \mathbf{D}. \quad (3.86)$$

In this case, Eq. (84) yields

$$g(\mathbf{r}) = \frac{1}{2} \mathbf{E} \cdot \mathbf{D} - \mathbf{E}_{\text{ext}} \cdot \mathbf{D} = \frac{\varepsilon}{2} E^2 - \varepsilon \mathbf{E} \cdot \mathbf{E}_{\text{ext}} \equiv \frac{\varepsilon}{2} (\mathbf{E} - \mathbf{E}_{\text{ext}})^2 + \text{const}, \quad (3.87)$$

where the constant may depend on the external field, but not on the resulting field distribution. As a sanity check, let us apply this result to a volume V well inside a long dielectric cylinder placed into a uniform external field \mathbf{E}_{ext} parallel to cylinder's axis. (Such orientation is important to ignore the geometric effects discussed in Sec. 3 – see, e.g., Fig. 6 and its discussion.) Then \mathbf{E} has to be uniform in the dominating part of the cylinder, so that Eq. (84) may be explicitly integrated over the volume, giving

$$\mathcal{G} = \frac{\varepsilon}{2} (\mathbf{E} - \mathbf{E}_{\text{ext}})^2 V + \text{const}. \quad (3.88)$$

The minimum of this function is achieved at the evidently correct result $\mathbf{E} = \mathbf{E}_{\text{ext}}$ - in contrast to the unphysical result $\mathbf{E} = 0$ (meaning electric field's expulsion from the volume) that we would get minimizing U .

3.6. Exercise problems

3.1.* Prove the following extension of Eq. (5):

$$\phi(\mathbf{r}) \approx \frac{1}{4\pi\varepsilon_0} \left(\frac{1}{r} Q + \frac{1}{r^3} \sum_{j=1}^3 r_j p_j + \frac{1}{2r^5} \sum_{j,j'=1}^3 r_j r_{j'} Q_{jj'} \right),$$

where Q is a scalar - the total charge of the system, p_j are the Cartesian components of a vector - system's dipole moment (6), and $Q_{jj'}$ are Cartesian components of a tensor - system's *quadrupole moment*:

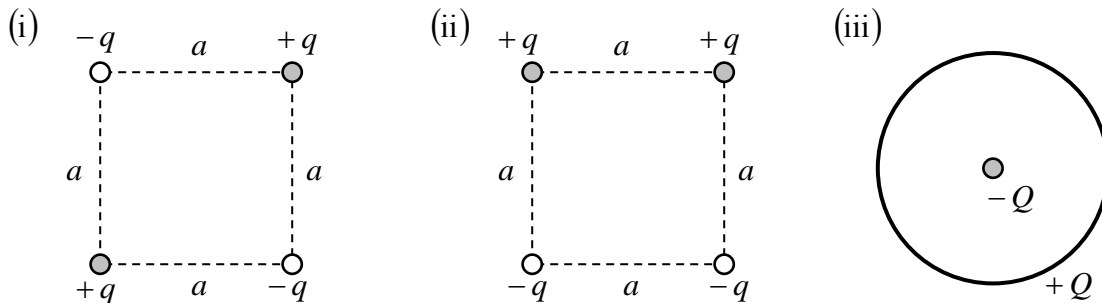
$$Q = \int \rho(\mathbf{r}') d^3 r', \quad p_j = \int \rho(\mathbf{r}') r'_j d^3 r', \quad Q_{jj'} = \int \rho(\mathbf{r}') (3r'_j r'_{j'} - r'^2 \delta_{jj'}) d^3 r'.$$

3.2. A plane, thin ring of radius R is charged with a constant linear density λ . Calculate the exact distribution electrostatic potential distribution along the symmetry axis of the ring, and prove that at large distances, $r \gg R$, it is indeed described by the multipole expansion spelled out in Problem 1.

3.3. Without carrying out an exact calculation, can you predict the spatial dependence of the interaction between various electric multipoles, including point charges (in this context, frequently called *monopoles*), dipoles, and quadrupoles? Based on these predictions, what is the functional dependence of the interaction between dumbbell-shaped diatomic molecules such as H_2 , N_2 , O_2 , etc., on the distance between them, if the distance is much larger than the molecular size?

3.4. In suitable reference frames, calculate the dipole and quadrupole moments of the following systems (see Figs. below):

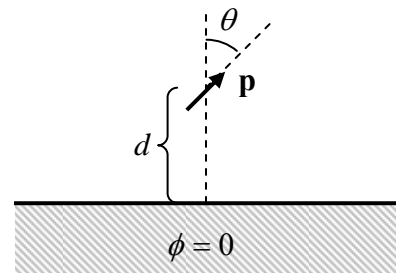
- (i) 4 point charges of the same magnitude, but alternating signs, placed in the corners of a square;
 (ii) a similar system, but with a pair charge sign alternation; and
 (iii) a point charge in the center of a thin ring carrying a similar but opposite charge, uniformly distributed along its circumference.



3.5. Two similar electric dipoles, of fixed magnitude p , located at fixed distance r from each other, are free to rotate, changing their directions. What stable equilibrium position(s) may they take as a result of their electrostatic interaction?

3.6. An electric dipole is located above an infinite conducting plane (see Fig. on the right). Calculate:

- the distribution of the induced charge in the conductor,
- the dipole-to-plane interaction energy, and
- the force and the torque acting on the dipole.

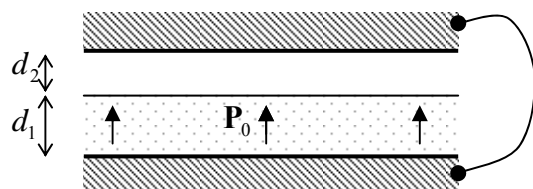


3.7. Use two different approaches to calculate the energy of interaction between a grounded conductor and an electric dipole \mathbf{p} , placed in the center of a spherical cavity of radius R , carved in the conductor.

3.8. A plane separating two parts of otherwise free space is densely and uniformly (with constant areal density n) filled with dipoles, with their dipole moments \mathbf{p} oriented in a direction normal to the plane.

- Calculate the boundary conditions for the electrostatic potential on both sides of the plane.
- Use the result of Task (i) to calculate the potential distribution created in space by a spherical surface, with radius R , densely and uniformly filled with radially-oriented dipoles.
- What condition that should be imposed on the dipole density n for your results to be qualitatively valid?

3.9. A plane capacitor, with zero voltage between its conducting plates (as may be fixed, e.g., with an external wire – see Fig. on the right), is partly filled with a material



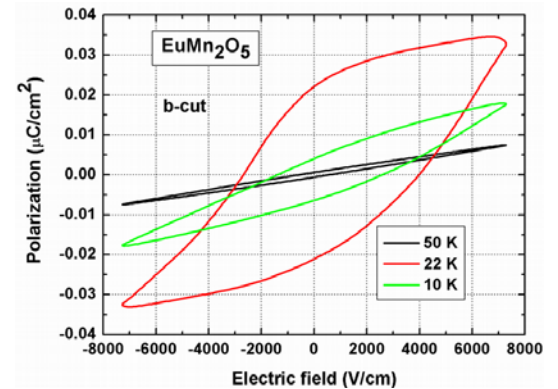
with spontaneous, constant polarization \mathbf{P}_0 .³² Find the distributions of the electric field, electric displacement, and the surface charge density of each plate.

3.10. A sphere of radius R is made of a material with a uniform, fixed polarization \mathbf{P}_0 .

- Calculate the electric field everywhere in space – both inside and outside the sphere.
- Explore the limit $R \rightarrow 0$, keeping $P_0 R^3 = \text{const}$, and compare the result with Eq. (25).

3.11. Discuss the physics of Eq. (3.85) of the lecture notes, in particular the physical nature of the potential energy U in a dipole medium. Apply your conclusion to a material with fixed (field-independent) polarization \mathbf{P}_0 , and calculate the electric field energy of the uniformly polarized sphere considered in the previous problem.

3.12. Experimental plots in Fig. on the right show that the polarization of EuMn_2O_5 , a typical ferroelectric/paraelectric material, becomes almost linear at 50 K. Use the plot to calculate (with an accuracy better than 10%) its dielectric constant ϵ_r at this temperature.



3.13. In two separate experiments, a thin, plane sheet of a linear dielectric with $\epsilon_r = \text{const}$ is placed into a uniform external electric field \mathbf{E}_0 :

- with sheet's surface parallel to the electric field, and
- the surface perpendicular to the field.

For each case, find the electric field \mathbf{E} , the electric displacement \mathbf{D} , and the polarization vector \mathbf{P} inside the dielectric (far from sheet's edges).

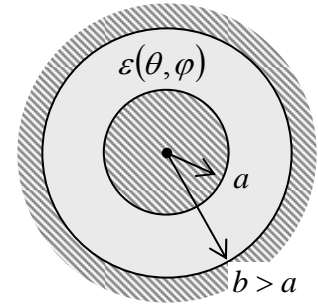
3.14. A point charge q is located at distance $r \gg R$ from the center of a uniform sphere of radius R , made of a uniform linear dielectric. In the first nonvanishing approximation in small parameter R/r , calculate the interaction force, and the energy of interaction between the sphere and the charge.

3.15. A fixed dipole \mathbf{p} is placed in the center of a spherical cavity of radius R , cut inside a uniform, linear dielectric. Calculate the electric field distribution everywhere in the system (both for $r < R$ and $r > R$).

Hint: You may start with the assumption that the field at $r > R$ has a distribution typical for a dipole (but be ready for surprises :-).

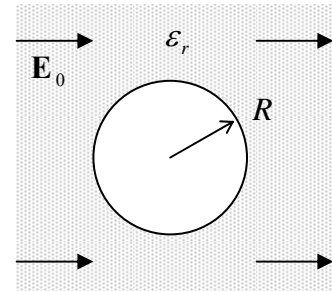
³² In electrical engineering, such materials (typically, synthetic polymers) are frequently called *electrets*. As an approximation, this condition may be applied to hard ferroelectrics, if the external or self-induced electric fields are not too high.

3.16. A spherical capacitor (see Fig. on the right) is filled with a linear dielectric whose permittivity ε depends on spherical angles θ and φ , but not on the distance r from system's center. Give an explicit expression for its capacitance C .



3.17. For each of the two capacitors shown in Fig. 3.7 of the lecture notes, calculate the electric forces (per unit area) on the boundaries of two uniform dielectrics, in terms of the electric fields.

3.18. A uniform electric field \mathbf{E}_0 has been created (by external sources) inside a uniform linear dielectric. Find the change of the electric field, created by cutting out a cavity in the shape of a round cylinder of radius R , with the axis perpendicular to the external field - see Fig. on the right.

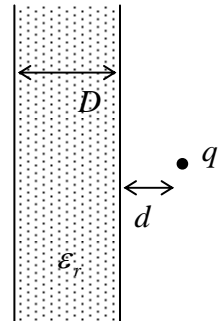


3.19. Small linear-dielectric particles of spherical shape are dispersed in free space with low concentration $n \ll 1/R^3$, where R is particle's radius. Calculate the average dielectric constant of such a medium. Compare the result with the apparent, but incomplete answer

$$\overline{\varepsilon_r} - 1 = (\varepsilon_r - 1)nV,$$

(where ε_r is the dielectric constant of particle's material and $V = (4\pi/3)R^3$ is its volume), and explain the origin of the difference.

3.20.* Calculate the spatial distribution of the electric potential induced by a point charge q is placed at distance d from a very wide parallel plate, of thickness D , made of a linear dielectric – see Fig. on the right.



Chapter 4. DC Currents

In this chapter I discuss the laws governing the distribution of constant (“dc”) currents inside conducting media, with a focus on the linear (“Ohmic”) conductivity. In most cases, the partial differential equation governing the distribution may be reduced to the same Laplace and Poisson equations whose solution methods have been discussed in detail in Chapter 2. Due to this fact, this chapter is rather short.

4.1. Continuity equation and the Kirchhoff laws

Until this point, our discussion of conductors has been limited to the cases when they are separated with *insulators* (meaning either vacuum or dielectric media) preventing any continuous motion of charges from one conductor to another, even if there is a voltage difference (and hence electric field) between them – see Fig. 1a.

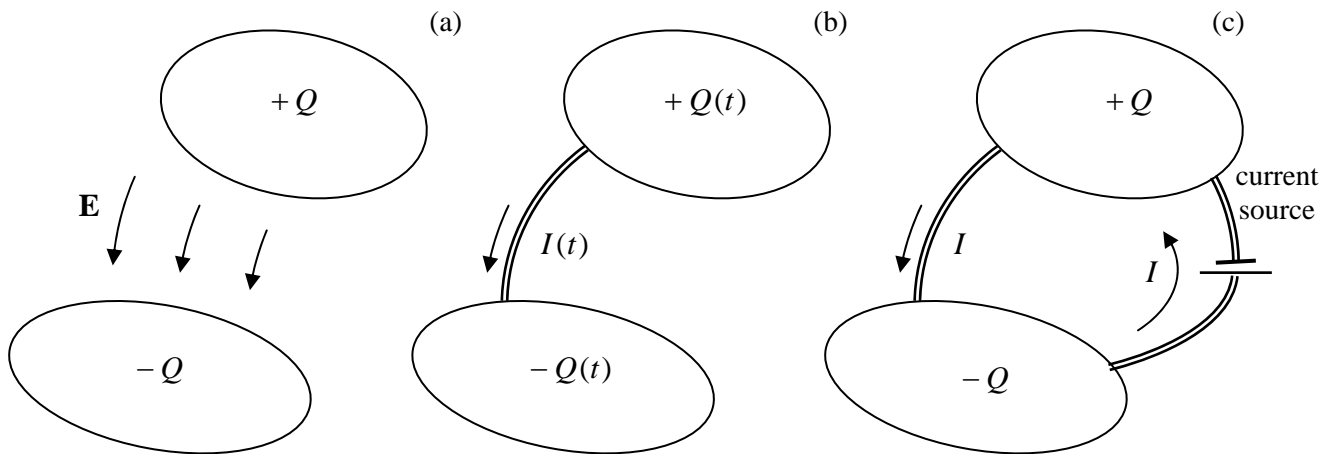


Fig. 4.1. Two oppositely charged conductors: (a) at the electrostatic situation, (b) at charge relaxation through an additional narrow conductor (“wire”), and (c) a system sustaining dc current in the wire.

Now let us connect two conductors galvanically, say with a *wire* – a thin, elongated conductor (Fig. 1b). Then the electric field causes the motion of charges in the wire - from a conductor with a higher electrostatic potential toward that with a lower potential, until the potentials equilibrate. Such process is called *charge relaxation*. The main equation governing this process may be obtained from the experimental fact (already mentioned in Sec. 1.1) that electric charges cannot appear or disappear (though opposite charges may recombine with the conservation of the net charge.) As a result the change of charge Q in one conductor may change only due to the current I through the wire:¹

$$\frac{dQ}{dt} = -I. \quad (4.1)$$

¹ Just as a (hopefully, unnecessary :-)) reminder, in the SI units the current is measured in amperes (A). In the legal metrology, the ampere (rather than the coulomb, which is defined as $1\text{C} = 1\text{A} \times 1\text{s}$) is a primary unit. I will mention its formal definition in the next chapter. In the Gaussian units, Eq. (1) remains the same, so that the current’s unit is the so-called *statampere* - defined as statcoulomb per second.

Let us express this law in a differential form, introducing the notion of *current density* vector $\mathbf{j}(\mathbf{r})$. This vector may be defined via the following relation for current dI crossing an elementary area dA (Fig. 2)

$$dI = j dA \cos \theta = (j \cos \theta) dA = j_n dA, \quad (4.2)$$

where θ is the angle between the normal to the surface and the carrier motion direction (which is taken for the direction of vector \mathbf{j}).

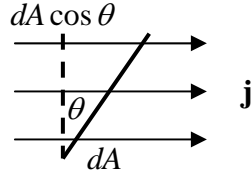


Fig. 4.2. Current density vector.

With that definition, Eq. (1) may be re-written as

$$\frac{d}{dt} \int_V \rho d^3 r = - \oint_S j_n d^2 r, \quad (4.3)$$

where V is an arbitrary stationary volume limited by the closed surface S . Applying to this volume the same divergence theorem as was repeatedly used in previous chapters, we get

$$\int_V \left[\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{j} \right] d^3 r = 0. \quad (4.4)$$

Since volume V is arbitrary, this equation may be true only if

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{j} = 0. \quad (4.5)$$

Continuity
equation

This is the fundamental *continuity equation* - which is true even for the time-dependent phenomena.²

The charge relaxation is of course a dynamic, time-dependent process. However, electric currents may also exist in stationary situations, when a current source, for example a *battery*, replenishes the conductor charges and hence sustains currents at a certain time-independent level – see Fig. 1c. (As we will see below, in most cases this process requires a persistent replenishment of the electrostatic energy from either a source or storage of energy of a different kind – say, the chemical energy of the battery.) Let us discuss the laws governing the distribution of such *dc currents*. In this case ($\partial/\partial t = 0$), Eq. (5) reduces to a very simple equation

$$\nabla \cdot \mathbf{j} = 0. \quad (4.6)$$

This equation acquires an even a simpler form in the particular but important case of *electric circuits* (Fig. 3), the systems may be presented as an electric connection of components of two types:

² Similar differential relations are valid for the density of any conserved quantity, for example for mass in classical fluid dynamics (see, e.g., CM Sec. 8.2), and for the probability in statistical physics (SM Sec. 5.6) and quantum mechanics (QM Sec. 1.4).

- (i) small-size (*lumped*) *circuit elements* (also called “two-terminal devices”), meaning a passive resistor, a current source, etc. – generally, any black box with two wires sticking out of it, and
- (ii) *perfectly conducting wires*, with negligible voltage drop along them, that are galvanically connected at certain points, called *nodes* (or “junctions”).

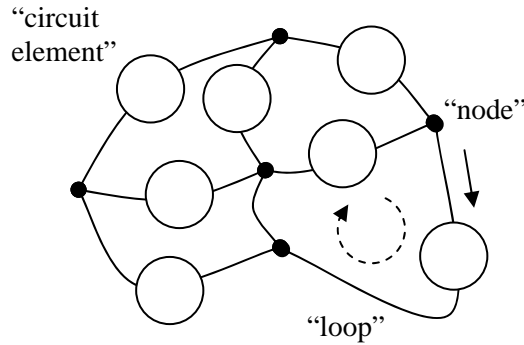


Fig. 4.3. Typical system obeying the Kirchhoff laws.

In the standard circuit theory, the electric charges of the nodes are considered negligible, and we may integrate Eq. (6) over the closed surface drawn around any node to get

$$\sum_j I_j = 0, \quad (4.7a)$$

where the summation is over all the wires (numbered with index j) connected in the node. On the other hand, according to its definition (2.25), voltage drop V_k across each circuit element may be presented as the difference of potentials of the adjacent nodes, $V_k = \phi_k - \phi_{k-1}$. Summing such differences around any closed loop of the circuit (Fig. 3), we get all terms cancelled, so that

$$\sum_k V_k = 0. \quad (4.7b)$$

These relations are called, respectively, the 1^{st} and 2^{nd} *Kirchhoff laws* - or sometimes the *node rule* and the *loop rule*. They may seem elementary, and the genuine power of the Kirchhoff approach is in the fact a set of Eqs. (7), covering every node and every circuit element of the system, gives a system of equations sufficient for the calculation of all currents and voltages in it - provided that the relation between current and voltage is known for each circuit element.

It is almost evident that in the absence of current sources, the system of equations (7) has only a trivial solution: $I_j = 0$, $V_k = 0$ - with the exotic exception of superconductivity, to be discussed in Sec. 6.3. The current sources, that allow non-vanishing current flows, may be described by their *electromotive forces* (*e.m.f.*) \mathcal{V}_k , having the dimensionality of voltage, which have to be taken into account in the corresponding terms V_k of sum (7b). Let me hope that the reader has some experience of using Eqs. (7) for the analysis of simple circuits – say consisting of several resistors and dc batteries – so I may save time on a discussion of these simple problems.

4.2. The Ohm law

As was mentioned above, the relations spelled out in Sec. 1 are sufficient for forming a closed system of equations for finding currents and electric field in a system only if they are complemented

with *material equations* relating scalars I and V in each circuit element, i.e. vectors \mathbf{j} and \mathbf{E} in each point of the medium of such an element. The simplest, and most frequently met relation of this kind is the famous *Ohm law* whose differential form is

$$\mathbf{j} = \sigma \mathbf{E}, \quad (4.8)$$

where σ is a constant called *conductivity*.³ Though this is not a fundamental relation, and is approximate for any conducting media, we can argue that if:

- (i) there is no current at $\mathbf{E} = 0$ (mind superconductors!),
- (ii) the medium is isotropic or almost isotropic (a notable exception: some organic conductors),
- (iii) the mean free path l of current carriers is much smaller than the characteristic scale a of the spatial variations of \mathbf{j} and \mathbf{E} ,

then the Ohm law may be viewed as a result of the Taylor expansion of the local relation $\mathbf{j}(\mathbf{E})$ in relatively small fields, and thus is very common.

Table 1 gives the experimental values of dc conductivity for some practically important (or just representative) materials. The reader can see that the range of its values is very broad, covering more than 30 orders of magnitude, even without going to such extremes as very pure metallic crystals at very low temperatures, where σ may reach $\sim 10^{12}$ S/m.

Table 4.1. Ohmic conductivities for some representative (or practically important) materials at 20°C.

Material	σ (S/m)
Teflon ($[\text{C}_2\text{F}_4]_n$)	10^{-22} - 10^{-24}
Silicon dioxide	10^{-16} - 10^{-19}
Various glasses	10^{-10} - 10^{-14}
Deionized water	$\sim 10^{-6}$
Sea water	5
Silicon n -doped to 10^{16}cm^{-1}	2.5×10^2
Silicon n -doped to 10^{19}cm^{-1}	1.6×10^4
Silicon p -doped to 10^{19}cm^{-1}	1.1×10^4
Nichrome (alloy 80% Ni + 20% Cr)	0.9×10^6
Aluminum	3.8×10^7
Copper	6.0×10^7
Zinc crystal along a -axis	1.65×10^7
Zinc crystal along c -axis	1.72×10^7

³ In SI units, the conductivity is measured in siemens per meter, where one siemens (S) is the reciprocal of one ohm: $1 \text{ S} \equiv (1 \Omega)^{-1} \equiv 1 \text{ A} / 1 \text{ V}$. The constant reciprocal to conductivity, $1/\sigma$, is called *resistivity*, and is commonly denoted by letter ρ . I will, however, try to avoid using this notion, because I am already overusing this letter.

In order to get some feeling what do these values mean, let us consider a very simple system (Fig. 4): a plane capacitor of area $A \gg d^2$, filled with a material that has not only a dielectric constant ϵ_r , but also some Ohmic conductivity σ , with much more conductive plate electrodes.

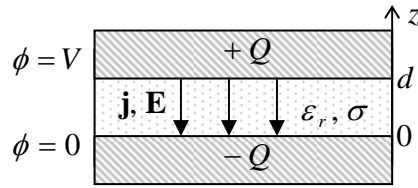


Fig. 4.4. “Leaky” plane capacitor.

Assuming that these properties are compatible with each other,⁴ we may assume that the distribution of electric potential (not too close to the capacitor edges) still obeys Eq. (2.39), so that the electric field is vertical and uniform, with $E = V/d$. Then, according to Eq. (6) the current density is also uniform, $j = \sigma E = \sigma V/d$. From here, the total current between the plates is

$$I = jA = \sigma EA = \sigma \frac{V}{d} A. \quad (4.9)$$

On the other hand, from Eqs. (2.26) and (3.45), the instant value of plate charge is $Q = C_m V = (\epsilon_r \epsilon_0 A/d)V$. Plugging these relations into Eq. (1), we see that the speed of charge (and voltage) relaxation does not depend on the geometric parameters A and d :

$$\frac{dV}{dt} = -\frac{V}{\tau_r}, \quad \tau_r \equiv \frac{\epsilon_r \epsilon_0}{\sigma}, \quad (4.10)$$

where parameter τ_r has the sense of the *relaxation time constant*. As we know (see Table 3.1), for most practical materials the dielectric constant is within one order of magnitude from 10, so that the nominator of Eq. (10) is of the order of 10^{-10} . As a result, according to Table 1, the charge relaxation time ranges from $\sim 10^{14}$ s (more than a million years!) for best insulators like teflon, to $\sim 10^{-18}$ s for the least resistive metals.

What is the physics behind these values of σ and why, for some materials, Table 1 gives them with such a large uncertainty? If charge carriers move as classical particles (e.g., in plasmas or non-degenerate semiconductors), a reasonable description of conductivity is given by the famous *Drude formula*.⁵ In his picture, due to weak electric field, the charge carriers are accelerated in its direction (possibly on the top of their random motion in all directions, i.e. with a vanishing average velocity vector):

$$\frac{d\mathbf{v}}{dt} = \frac{q}{m} \mathbf{E}, \quad (4.11)$$

and as a result their velocity acquires an the average value

$$\langle \mathbf{v} \rangle = \frac{d\mathbf{v}}{dt} \tau = \frac{q}{m} \mathbf{E} \tau, \quad (4.12)$$

⁴ As will be discussed in Chapter 6, such simple analysis is only valid if σ is not too high.

⁵ It was suggested by P. Drude in 1900.

where the phenomenological parameter $\tau = l/v$ (not to be confused with τ_r !) may be understood as the effective average time between carrier scattering events. From here, the current density:

$$\mathbf{j} = qn\langle\mathbf{v}\rangle = \frac{q^2 n \tau}{m} \mathbf{E}, \quad \text{i.e. } \sigma = \frac{q^2 n \tau}{m}. \quad (4.13a)$$

(Notice the independence of σ of the carrier charge sign.) Another form of the same result, more popular in the physics of semiconductors, is

$$\sigma = q^2 n \mu, \quad \text{with } \mu = \frac{\tau}{m}, \quad (4.13b)$$

Two
versions
of the
Drude
formula

where parameter μ , defined by relation $\langle\mathbf{v}\rangle \equiv \mu\mathbf{E}$, is called the charge carrier *mobility*.

Most good conductors (e.g., metals) are essentially degenerate Fermi gases (or liquids), in which the average thermal energy of a particle, $k_B T$ is much lower than the Fermi energy ϵ_F . In this case, a quantum theory is needed for the calculation of σ . Such theory was developed by the quantum physics' godfather A. Sommerfeld in 1927 (and is sometimes called the *Drude-Sommerfeld model*). I have no time to discuss it in this course,⁶ and here I will only notice that for an ideal, isotropic Fermi gas the result is reduced to Eq. (13), with a certain effective value of τ , so it may be used for estimates of σ , with due respect to the quantum theory of scattering. In a typical metal, n is very high ($\sim 10^{23} \text{ cm}^{-3}$) and is fixed by the atomic structure, so that the sample quality may only affect σ via the scattering time τ .

At room temperature, the scattering of electrons by thermally-excited lattice vibrations (*phonons*) dominates, so that τ and σ are high but finite, and do not change much from one sample to another. (Hence, the more accurate values given for metals in Table 1.) On the other hand, at $T \rightarrow 0$, a perfect crystal should not exhibit scattering at all, and conductivity should be infinite. In practice, this is never true (for example, due to electron scattering from imperfect boundaries of finite-size samples), and the effective conductivity σ is infinite (or practically infinite, at least above the measurable value $\sim 10^{20} \text{ S/m}$) only in superconductors.⁷

On the other hand, the conductivity of quasi-insulators (including deionized water) and semiconductors depends mostly of the carrier density n that is much lower than in metals. From the point of view of quantum mechanics, this happens because the ground-state eigenenergies of charge carriers are localized within an atom (or molecule), and separated from excited states, with space-extended wavefunctions, by a large energy gap (called *bandgap*). For example, in SiO_2 the bandgap approaches 9 eV, equivalent to $\sim 4,000 \text{ K}$. This is why, even at room temperatures the density of thermally-excited free charge carriers in good insulators is negligible. In these materials, n is determined by impurities and vacancies, and may depend on a particular chemical synthesis or other fabrication technology, rather than on fundamental properties of the material. (On the contrary, the carrier mobility μ in these materials is almost technology-independent.)

The practical importance of the technology may be illustrated on the following example. In cells of the so-called *floating-gate memories*, in particular the *flash memories*, which currently dominate the nonvolatile digital memory technology, data bits are stored as small electric charges ($Q \sim 10^{-16} \text{ C}$) of

⁶ For such a discussion see, e.g., SM Sec. 6.3.

⁷ Electrodynamical properties of superconductors are so interesting (and important) that I will discuss them in more detail in Chapter 6.

highly doped silicon islands (so-called *floating gates*) separated from the rest of the integrated circuit with a ~ 10 -nm-thick layer of silicon dioxide, SiO_2 . Such layers are fabricated by high-temperature oxidation of virtually perfect silicon crystals. The conductivity of the resulting high-quality (though amorphous) material is so low, $\sigma \sim 10^{-19}$ S/m, that the relaxation time τ_r , defined by Eq. (10), is well above 10 years – the industrial standard for data retention in non-volatile memories. In order to appreciate how good this technology is, the cited value should be compared with the typical conductivity $\sigma \sim 10^{-16}$ S/m of the usual, bulk SiO_2 ceramics.⁸

4.3. Boundary problems

For an Ohmic conducting media, we may combine Eqs. (6) and (8) the following differential equation

$$\nabla \cdot (\sigma \nabla \phi) = 0. \quad (4.14)$$

For a uniform conductor ($\sigma = \text{const}$), Eq. (14) is reduced to the Laplace equation for the electrostatic potential ϕ . As we already know from Chapters 2 and 3, its solution depends on the boundary conditions. These conditions depend on the interface type.

(i) Conductor-conductor interface. Applying the continuity equation (6) to a Gauss-type pillbox at the interface of two different conductors (Fig. 5), we get

$$(j_n)_1 = (j_n)_2, \quad (4.15)$$

so that if the Ohm law is valid inside each medium, then

$$\sigma_1 \frac{\partial \phi_1}{\partial n} = \sigma_2 \frac{\partial \phi_2}{\partial n}. \quad (4.16)$$

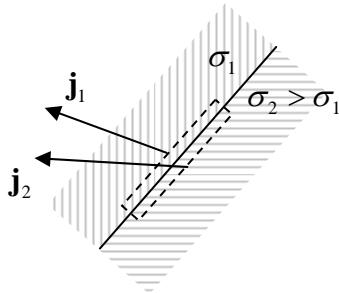


Fig. 4.5. DC current “refraction” at the interface between two different conductors.

Also, since the electric field should be finite, its potential ϕ has to be continuous across the interface - the condition that may also be written as

⁸ Unfortunately, these notes are not an appropriate platform to discuss details of the floating-gate memory technology. However, I think that every educated physicist should know its basics, because such memories are presently the driver of all semiconductor integrated circuit technology development, and hence of the whole information technology progress. Perhaps the best available book is J. Brewer and M. Gill (eds.), *Nonvolatile Memory Technologies with Emphasis on Flash*, IEEE, 2008.

$$\frac{\partial \phi_1}{\partial \tau} = \frac{\partial \phi_2}{\partial \tau}. \quad (4.17)$$

Both these conditions (and hence the solutions of the boundary problems using them) are similar to those for the interface between two dielectrics – cf. Eqs. (3.46)-(3.47).

Note that using the Ohm law, Eq. (17) may be rewritten as

$$\frac{1}{\sigma_1}(j_\tau)_1 = \frac{1}{\sigma_2}(j_\tau)_2. \quad (4.18)$$

Comparing it with Eq. (15) we see that, generally, the current density magnitude changes at the interface: $j_1 \neq j_2$. It is also curious that if $\sigma_1 \neq \sigma_2$, the current line slope changes at the interface (Fig. 4), qualitatively to the refraction of light rays in optics – see Chapter 7.

(ii) Conductor-electrode interface. The definition of an *electrode*, or a “perfect conductor”, is a medium with $\sigma \rightarrow \infty$. Then, at fixed current density at the interface, the electric field in the electrode tends to zero, and hence it may be described by equation

$$\phi = \phi_j = \text{const}, \quad (4.19)$$

where constants ϕ_j may be different for different electrodes (numbered with index j). Note that with such boundary conditions the Laplace boundary problem becomes exactly the same as in electrostatics – see Eq. (2.35) – and hence we can use all the methods (and some solutions :-) of Chapter 2 for finding dc current distribution.

(iii) Conductor-insulator interface. For the description of an insulator, we can use $\sigma = 0$, so that Eq. (16) yields the following boundary condition,

$$\frac{\partial \phi}{\partial n} = 0, \quad (4.20)$$

for the potential derivative *inside the conductor*. From the Ohm law we see that this is just the very natural requirement for the dc current not to flow into an insulator.

Now, note that this condition makes the Laplace problem inside the conductor completely well-defined, and independent on the potential distribution in the adjacent insulator. On the contrary, due to the continuity of the electrostatic potential at the border, its distribution in the insulator has to follow that inside the conductor. Let us discuss this conceptual issue on the following (apparently, trivial) example: dc current in a long wire with a constant cross-section area A . The reader certainly knows the answer:

$$I = \frac{V}{R}, \quad \text{where } R \equiv \frac{V}{I} = \frac{l}{\sigma A}, \quad (4.21)$$

Uniform
wire's
resistance

where l is the wire length, and constant R is called the *resistance*.⁹ However, let us get this result formally from our theoretical framework. For the ideal geometry shown in Fig. 6a, this is easy to do. Here the potential evidently has a linear 1D distribution

⁹ The first of Eqs. (21) is essentially the integral form of the Ohm law (8), and is valid not only for a uniform wire, but for any Ohmic conductor with a geometry in which I and V may be clearly defined.

$$\phi = \text{const} - \frac{x}{l}V, \quad (4.22)$$

both in the conductor and the surrounding free space, with both boundary conditions (16) and (17) satisfied at the conductor-insulator interfaces, and condition (20) satisfied at the conductor-electrode interfaces. As a result, the electric field is constant and has only one component $E_x = V/l$, so that inside the conductor

$$j_x = \sigma E_x, \quad I = j_x A, \quad (4.23)$$

giving us the well-known Eq. (21).

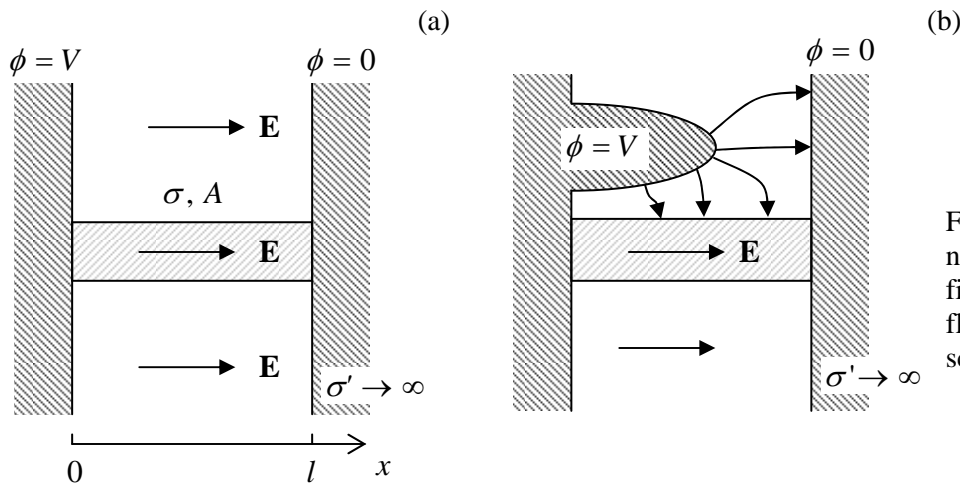


Fig. 4.6. (a) Trivial and (b) not-so-trivial problems of the field distribution at dc current flow. (For the latter case, schematically.)

However, what about the geometry shown in Fig. 6b? In this case the field distribution in the insulator is dramatically different, but according to boundary problem defined by Eqs. (14) and (20), inside the conductor the solution is exactly the same as it was in the former case. Now, the Laplace equation in the surrounding insulator has to be solved with the boundary values of the electrostatic potential, “dictated” by the distribution of the current (and hence potential) in the conductor.

Let us solve a problem in that this *conduction hierarchy* may be followed analytically to the very end. Consider an empty spherical cavity cut in a conductor with an initially uniform current flow with constant density $\mathbf{j}_0 = \mathbf{n}j_0$ (Fig. 7a). Following the conduction hierarchy, we have to solve the boundary problem in the conducting part of the system, i.e. outside the sphere ($r \geq R$), first. Since the problem is evidently axially-symmetric, we already know the general solution of the Laplace equation – see Eq. (2.172). Moreover, we know that in order to match the uniform field at $r \rightarrow \infty$, all coefficients a_l but one ($a_1 = -E_0 = -j_0/\sigma$) have to be zero, and that the boundary conditions at $r = R$ will give zero solutions for all coefficients b_l but one (b_1), so that

$$\phi = -\frac{j_0}{\sigma} r \cos \theta + \frac{b_1}{r^2} \cos \theta, \quad \text{for } r \geq R. \quad (4.24)$$

In order to find coefficient b_1 , we have to use the boundary condition (20) at $r = R$:

$$\left. \frac{\partial \phi}{\partial r} \right|_{r=R} = \left(-\frac{j_0}{\sigma} - \frac{2b_1}{R^3} \right) \cos \theta = 0. \quad (4.25)$$

This gives $b_1 = -j_0 R^3 / 2\sigma$, so that, finally,

$$\phi(r, \theta) = -\frac{j_0}{\sigma} \left(r + \frac{R^3}{2r^2} \right) \cos \theta. \quad (4.26)$$

(Note that this potential distribution corresponds to the dipole moment $\mathbf{p} = -\mathbf{E}_0 R^3 / 2$. It is easy to check that if the empty sphere was cut in a dielectric, the potential distribution outside the cavity would be similar, with $\mathbf{p} = -\mathbf{E}_0 R^3 (\epsilon_r - 1) / (\epsilon_r + 2)$. In the limit $\epsilon_r \rightarrow \infty$, these two results coincide, despite the rather different type of the problem: in the dielectric case, there is no current at all.)

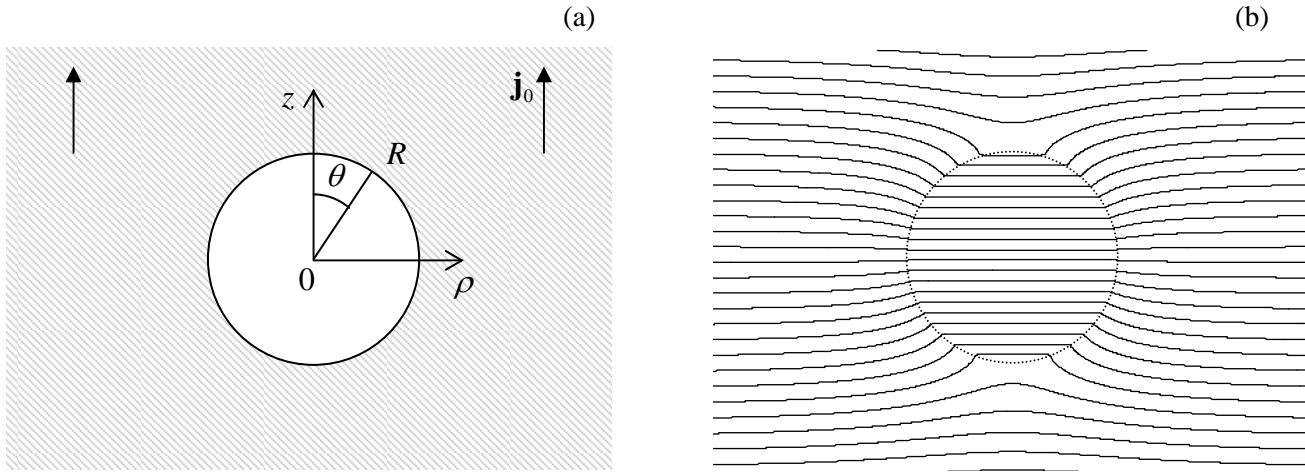


Fig. 4.7. Spherical cavity in a uniform conductor: (a) the problem's geometry, and (b) the equipotential surfaces, as given by Eq. (40) for $r > R$ and Eq. (42) for $r < R$.

Now, as the second step in the conductivity hierarchy, we may find the electrostatic potential distribution $\phi(r, \theta)$ in the insulator, in this particular case inside the cavity ($r \leq R$). It should also satisfy the Laplace equation with the boundary conditions at $r = R$, “dictated” by distribution (26):

$$\phi(R, \theta) = -\frac{3}{2} \frac{j_0}{\sigma} R \cos \theta. \quad (4.27)$$

We could again solve this problem by the formal variable separation (keeping in the general solution (2.172) only the term proportional to b_1 , that does not diverge at $r \rightarrow 0$), but if we notice that boundary condition (27) depends on just one Cartesian coordinate, $z = R \cos \theta$, the solution may be just guessed:

$$\phi(r, \theta) = -\frac{3}{2} \frac{j_0}{\sigma} z = -\frac{3}{2} \frac{j_0}{\sigma} r \cos \theta, \quad \text{at } r \leq R. \quad (4.28)$$

It evidently satisfies the Laplace equation and the boundary condition (27), and corresponds to a constant vertical electric field equal to $3j_0/2\sigma$ – see Fig. 6b.

The conductivity hierarchy says that static electrical fields and charges outside conductors (e.g., electric wires) do not affect currents flowing in the wires, and it is physically clear why. For example, if

a charge in vacuum is slowly moved close to a wire, it (in accordance with the linear superposition principle) will only induce an additional surface charge (see Chapter 2) that screens the external charge's field, without participating in (or disturbing) the current flow inside the conductor.

Besides the conceptual discussion, the two examples given above may be considered as a demonstration of the application of the first two methods described in Chapter 2 (the orthogonal coordinates (Fig. 5) and variable separation (Fig. 6)) to dc current distribution problems. Continuing this review of the methods we know, let us discuss the analog of the method of charge images. Let us consider the spherically-symmetric potential distribution of the electrostatic potential, similar to that given by Eq. (1.35):

$$\phi = \frac{c}{r}. \quad (4.29)$$

As we know from Chapter 1, this is a particular solution of the 3D Laplace equation at all points but $r = 0$, and hence is a legitimate solution in a current-carrying conductor as well. In vacuum, this distribution would correspond to a point charge $q = 4\pi\epsilon_0 c$; but what about the conductor? Calculating the corresponding electric field and current density,

$$\mathbf{E} = -\nabla\phi = \frac{c}{r^3}\mathbf{r}, \quad \mathbf{j} = \sigma\mathbf{E} = \sigma\frac{c}{r^3}\mathbf{r}, \quad (4.30)$$

we see that the total current flowing from the point in the origin through a sphere of an arbitrary radius r does not depend on the radius:

$$I = A j = 4\pi r^2 j = 4\pi\sigma c. \quad (4.31)$$

Plugging the resulting c into Eq. (29), we get

$$\phi = \frac{I}{4\pi\sigma r}. \quad (4.32)$$

Hence the Coulomb-type distribution of the electric potential in a conductor is possible (at least at some distance from the singular point $r = 0$), and describes dc current I flowing out of a small-size electrode - or *into* such a point, if coefficient c is negative. Such *current injection* may be readily implemented experimentally; think for example about an insulated wire with a small bare end, inserted into a poorly conducting soil – an important method in geophysical research.¹⁰

Now let the injection point \mathbf{r}' be close to a plane interface between the conductor and an insulator (Fig. 8). In this case, besides the Laplace equation, we should satisfy the boundary condition,

$$j_n = \sigma E_n = -\sigma \frac{\partial\phi}{\partial n} = 0. \quad (4.33)$$

It is clear that this can be done by replacing the insulator for a conductor with an additional current injection point, at the mirror image point \mathbf{r}'' . Note, however, that in contrast to the charge images, the sign of the imaginary current has to be *similar*, not opposite, to the initial one, so that the total electrostatic potential inside the conducting semi-space is

¹⁰ Such situations are even more natural in 2D situations, for example, think about a wire soldered, in a small spot, to a thin metallic foil. (Note that here the current density distribution law is different, $j \propto 1/r$ rather than $1/r^2$.)

$$\phi(\mathbf{r}) = \frac{I}{4\pi\sigma} \left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} + \frac{1}{|\mathbf{r} - \mathbf{r}''|} \right). \quad (4.34)$$

(Note that the image current's sign would be opposite if we discussed an interface between a conductor with a moderate conductivity and a perfect conductor (“electrode”) whose potential should be virtually constant.)

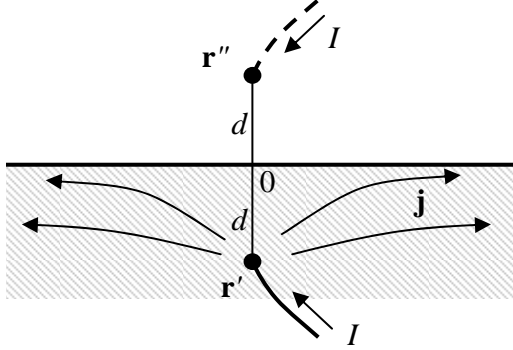


Fig. 4.8. Method of images at dc conduction.

This result may be readily used, for example, to calculate the current density at the conductor's surface, as a function of distance ρ from point 0 (the surface point closest to the current injection) – see Fig. 8. At the surface, Eq. (34) yields

$$\phi = \frac{I}{2\pi\sigma} \frac{1}{(\rho^2 + d^2)^{1/2}}, \quad (4.35)$$

so that the current density is independent of σ :

$$j_\rho = \sigma E_\rho = -\sigma \frac{\partial \phi}{\partial \rho} = \frac{I}{2\pi} \frac{\rho}{(\rho^2 + d^2)^{3/2}}. \quad (4.36)$$

Deviations from Eqs. (35) and (36), which are valid for a uniform medium, may be used to find and characterize conductance inhomogeneities, say, those due to mineral deposits in the Earth crust.¹¹

4.4. Dissipation power

Let me conclude this brief chapter with an ultra-short discussion of energy dissipation in conductors. In contrast to the electrostatics situations in insulators (vacuum or dielectrics), at dc conduction the electrostatic energy U is “dissipated” (i.e. transferred to heat) at a certain rate $\mathcal{P} \equiv -dU/dt$, called *dissipation power*.¹² This rate may be evaluated by calculating the power of electric field's work on a single moving charge:

¹¹ In practice, the current injection may be produced, due to electrochemical reactions, by an ore mass itself, so that one need only measure (and interpret :-) the resulting potential distribution - the so-called *self-potential method* - see, e.g., Sec. 6.1 in monograph by W. Telford *et al.*, *Applied Geophysics*, 2nd ed., Cambridge U. Press, 1990.

¹² Since the electric field and hence the electrostatic energy are time-independent, this means that the energy is replenished at the same rate from the current source(s).

$$\mathcal{P}_1 = \mathbf{F} \cdot \mathbf{v} = q\mathbf{E} \cdot \mathbf{v}. \quad (4.37)$$

After the summation over all charges, Eq. (37) gives us the dissipation power. If the charge density n is uniform, multiplying by it both parts of this equation, and taking into account that $qn\mathbf{v} = \mathbf{j}$, for the power dissipated in a unit volume we get the *Joule law*

General
Joule
law

$$p \equiv \frac{\mathcal{P}}{V} = \frac{\mathcal{P}_1 N}{V} = \mathcal{P}_1 n = q\mathbf{E} \cdot \mathbf{v} n = \mathbf{E} \cdot \mathbf{j}. \quad (4.38)$$

In the particular case of the Ohmic conductivity, this expression may be also rewritten in two other forms:

Joule law
for Ohmic
conductivity

$$p = \sigma E^2 = \frac{j^2}{\sigma}. \quad (4.39)$$

At dc conduction, the energy is permanently replenished by a flow of power from the current source(s).

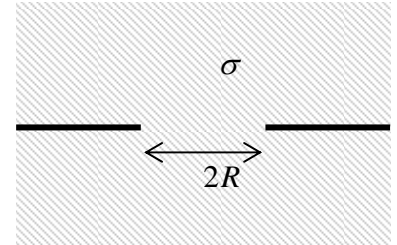
With our electrostatics background, it is straightforward (and hence left for reader's exercise) to prove that the dc current distribution in a uniform Ohmic conductor, at a fixed voltage applied at its borders, corresponds to the minimum of the total dissipation power

$$\mathcal{P} = \sigma \int_V E^2 d^3 r. \quad (4.40)$$

4.5. Exercise problems

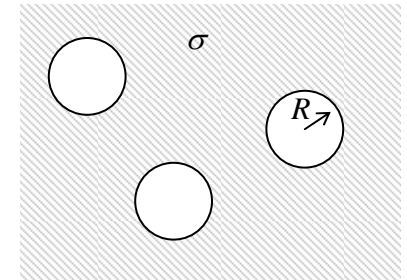
4.1. Find the resistance between two large conductors separated with a very thin, plane, insulating partition, with a circular hole of radius R in it – see Fig. on the right.

Hint: You may like to use the degenerate ellipsoidal coordinates that had been used in Sec. 2.4 to find the self-capacitance of a round disk in vacuum.



4.2. Calculate the effective (average) conductivity σ_{ef} of a medium with many empty spherical cavities of radius R , carved at random points in a uniform Ohmic conductor (see Fig. on the right), in the limit of low density $n \ll R^{-3}$ of the spheres.

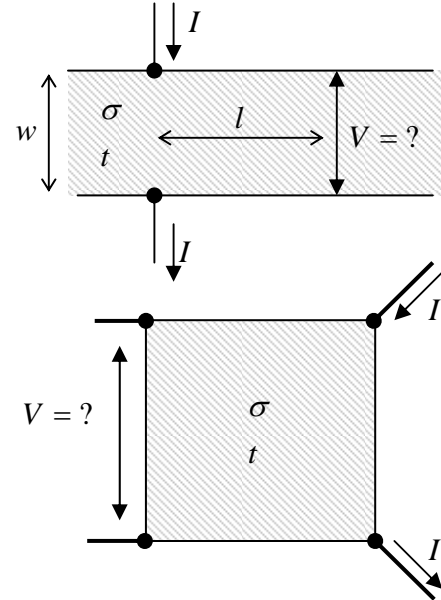
Hint: Try to use the analogy with a dipole media (Sec. 3.2).



4.3. In two separate experiments, a narrow gap, of irregular width, between two close metallic electrodes is filled with some material - in the first case, with a uniform linear insulator with an electric permittivity ϵ , and in the second case, with a uniform conducting material with an Ohmic conductivity σ . Neglecting the fringe effects, calculate the relation between the mutual capacitance C between the electrodes (in the first case) and the dc resistance R between them (in the second case).

4.4. Calculate the voltage drop V across a uniform, wide resistive slab of thickness t , at distance l from the points of injection/ejection of dc current I that is passed across the slab - see Fig. on the right.

Hint: Try to use the dc current analog of the charge image method.



4.5. Find the voltage drop V between two corners of a square cut from a uniform, resistive sheet of thickness t , induced by dc current I that is passed between its two other corners - see Fig. on the right.

4.6. Calculate the distribution of dc current density in a thin, round, uniform resistive disk, if the current is inserted into a point at its rim, and picked up at the center.

4.7.* The simplest model of a vacuum diode consists of two plane, parallel metallic electrodes of area A , separated by a gap of thickness $d \ll A^{1/2}$: a “cathode” which emits electrons to vacuum, and an “anode” which absorbs the electrons arriving at its surface. Calculate the dc I - V curve of the diode, i.e. the stationary relation between current I flowing between the electrodes and voltage V applied between them, using the following simplifying assumptions:

- (i) due to the effect of the negative space charge of the emitted electrons, current I is much smaller than the emission ability of the cathode,
- (ii) the initial velocity of the emitted electrons is negligible, and
- (iii) the direct Coulomb interaction of electrons (besides the space charge effect) is negligible.

4.8.* Calculate the space-charge-limited current in a system with the same geometry, and using the same assumptions as in the previous problem, besides assuming now that the emitted charge carriers move not ballistically, but in accordance with the Ohm law, with the conductivity given by Eq. (4.13): $\sigma = q^2 \mu n$, with constant mobility μ .

Hint: In order to get a realistic result, assume that the medium in which the carriers move¹³ has a certain dielectric constant ϵ_r .

4.9. Prove that the distribution of dc currents in a uniform Ohmic conductor, at fixed voltage applied at its boundaries, corresponds to the minimum of the total power dissipation (“Joule heat”).

¹³ As was mentioned in Sec. 4.2 of the lecture notes, the assumption of constant (charge-density-independent) mobility is most suitable for semiconductors.

Chapter 5. Magnetism

Despite the fact that we are now starting to discuss a completely new type of electromagnetic interactions, its coverage (for the stationary case) will take just one chapter, because we will be able to recycle many ideas and methods of electrostatics, though with a twist or two.

5.1. Magnetic interaction of currents

DC currents in conductors usually leave them *electroneutral*, $\rho(\mathbf{r}) = 0$, with a very good precision, because any virtual misbalance of positive and negative charge density results in extremely strong Coulomb forces that restore their balance by an additional shift of free carriers.¹ This is why let us start the discussion of magnetic interactions from the simplest case of two spatially-separated, current-carrying, electroneutral conductors (Fig. 1).

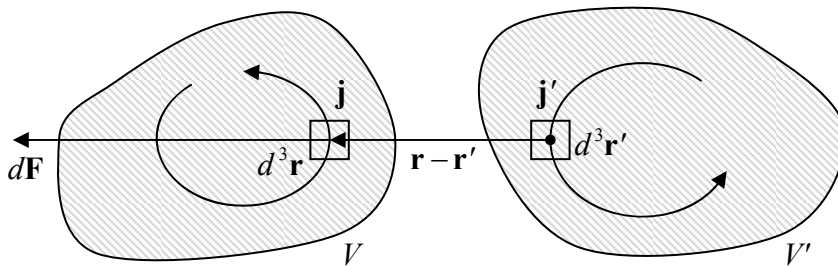


Fig. 5.1. Magnetic interaction of two currents.

According to the Coulomb law, there should be no force between them. However, several experiments carried out in the early 1820s² proved that such non-Coulomb forces do exist, and are the manifestation of another, *magnetic* interactions between the currents. In the contemporary used in this course, their results may be summarized with one formula, in SI units expressed as:³

$$\mathbf{F} = -\frac{\mu_0}{4\pi} \int_V d^3r \int_{V'} d^3r' (\mathbf{j}(\mathbf{r}) \cdot \mathbf{j}'(\mathbf{r}')) \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3}. \quad (5.1)$$

Here coefficient $\mu_0/4\pi$ (where μ_0 is called either the *magnetic constant* or the *free space permeability*), by definition, equals *exactly* 10^{-7} SI units, thus relating the electric current (and hence electric charge) definition to that of force – see below.

Note that the Coulomb law (1.1), with the account of the linear superposition principle, may be presented in a very similar form:

¹ The most important case when the electroneutrality does not hold is the motion of electrons in vacuum. In this case, magnetic forces coexist with (typically, stronger) electrostatic forces – see Eq. (3) below and its discussion. In some semiconductor devices, local violations of electroneutrality also play an important role.

² Most notably, by H. C. Ørsted, J.-B. Biot and F. Savart, and A.-M. Ampère.

³ In the Gaussian units, coefficient $\mu_0/4\pi$ is replaced with $1/c^2$ (i.e., implicitly with $\mu_0\epsilon_0$) where c is the speed of light, in modern metrology considered *exactly* known – see, e.g., appendix CA: *Selected Physical Constants*.

$$\mathbf{F} = \frac{1}{4\pi\epsilon_0} \int_V d^3r \int_{V'} d^3r' \rho(\mathbf{r}) \rho'(\mathbf{r}') \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3}. \quad (5.2)$$

Besides the different coefficient and sign, the “only” difference of Eq. (1) from Eq. (2) is the scalar product of current densities, evidently necessary because of the vector character of the current density. We will see that this difference will bring certain complications in applying the electrostatics approaches, discussed in the previous chapters, to magnetostatics.

Before going to their discussion, let us have one more glance at the coefficients in Eqs. (1) and (2). To compare them, let us consider two objects with uncompensated charge distributions $\rho(\mathbf{r})$ and $\rho'(\mathbf{r})$, each moving parallel to each other as a whole certain velocities \mathbf{v} and \mathbf{v}' , as measured in an inertial “lab” frame. In this case, $\mathbf{j}(\mathbf{r}) = \rho(\mathbf{r})\mathbf{v}$, $\mathbf{j}(\mathbf{r}) \cdot \mathbf{j}'(\mathbf{r}) = \rho(\mathbf{r})\rho'(\mathbf{r})\mathbf{v}\mathbf{v}'$, and the integrals in Eqs. (1) and (2) become functionally similar, and differ only by the factor

$$\frac{F_{\text{magnetic}}}{F_{\text{electric}}} = -\frac{\mu_0 \mathbf{v}\mathbf{v}'}{4\pi} / \frac{1}{4\pi\epsilon_0} = -\frac{\mathbf{v}\mathbf{v}'}{c^2}. \quad (5.3)$$

(This expression hold in any consistent system of units.) We immediately see that magnetism is an essentially relativistic phenomenon, very weak in comparison with the electrostatic interaction at the human scale velocities, $v \ll c$, and may dominate only if the latter interaction vanishes – as it does in electroneutral systems.⁴

Also, Eq. (3) points at an interesting paradox. Consider two electron beams moving parallel to each other, with the same velocity v with respect to a lab reference frame. Then, according to Eq. (3), the net force of their total (electric and magnetic) interaction is proportional to $(1 - v^2/c^2)$, and tends to zero in the limit $v \rightarrow c$. However, in the reference frame moving together with electrons, they are not moving at all, i.e. $v = 0$. Hence, from the point of view of such a moving observer, the electron beams should interact only electrostatically, with a repulsive force independent of velocity v . Historically, this had been one of several paradoxes that led to the development of the special relativity; its resolution will be discussed in Chapter 9, devoted to this theory.

Returning to Eq. (1), in some simple cases, the double integration in it may be carried out analytically. First of all, let us simplify this expression for the case of two thin, long conductors (wires) separated by a distance much larger than their thickness. In this case we may integrate the products $\mathbf{j}d^3r$ and $\mathbf{j}'d^3r'$ over wires' cross-sections first, neglecting the corresponding change of $(\mathbf{r} - \mathbf{r}')$. Since the integrals of the current density over the cross-sections of the wire are just the currents I and I' in the wires, and cannot change along their lengths (correspondingly, l and l'), they may be taken out of the remaining integrals, reducing Eq. (1) to

$$\mathbf{F} = -\frac{\mu_0 II'}{4\pi} \oint_l \oint_{l'} (d\mathbf{r} \cdot d\mathbf{r}') \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3}. \quad (5.4)$$

⁴ The discovery and initial studies of such a subtle, relativistic phenomenon as magnetism in the early 19th century was much facilitated by the relative abundance of natural *ferromagnets*, materials with spontaneous magnetic polarization, whose strong magnetic field may be traced back to relativistic effects (such as spin) in atoms. (The electrostatic analogs of such materials, *electrets*, are much more rare.) I will briefly discuss the ferromagnetism in Sec. 5 below.

As the simplest example, consider two straight, parallel wires (Fig. 2), separated by distance d , with length $l \gg \rho$. In this case, due to symmetry, the vector of magnetic interaction force has to:

- (i) lay in the same plane as the currents, and
- (ii) be perpendicular to the wires – see Fig. 2.

Hence we can limit our calculations to just one component of the force. Using the fact that with the coordinate choice shown in Fig. 2, $d\mathbf{r} \cdot d\mathbf{r}' = dx dx'$, we get

$$F = -\frac{\mu_0 I I'}{4\pi} \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dx' \frac{\sin \theta}{d^2 + (x - x')^2} = -\frac{\mu_0 I I'}{4\pi} \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dx' \frac{d}{[d^2 + (x - x')^2]^{3/2}}. \quad (5.5)$$

Introducing, instead of x' , a new, dimensionless variable $\xi \equiv (x - x')/\rho$, we may reduce the internal integral to a table integral which we have already met in this course:

$$F = -\frac{\mu_0 I I'}{4\pi d} \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} \frac{d\xi}{(1 + \xi^2)^{3/2}} = -\frac{\mu_0 I I'}{2\pi d} \int_{-\infty}^{+\infty} dx. \quad (5.6)$$

The integral over x is formally diverging, but this means merely that the interaction force *per unit length* of the wires is constant:

$$\frac{F}{l} = -\frac{\mu_0 I I'}{2\pi d}. \quad (5.7)$$

Note that the force drops rather slowly (only as $1/d$) as the distance d between the wires is increased, and is *attractive* (rather than repulsive as in the Coulomb law) if the currents are of the same sign.

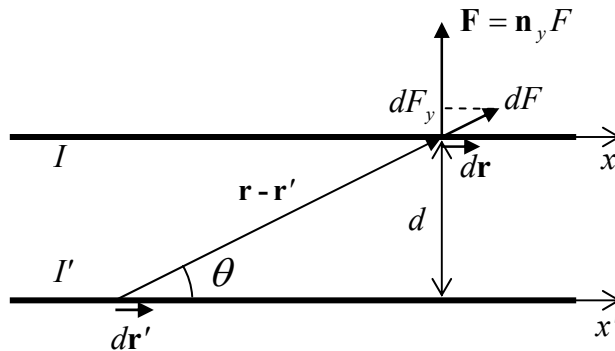


Fig. 5.2. Magnetic force between two straight parallel currents.

This is an important result,⁵ but again, the problems solvable so simply are few and far between, and it is intuitively clear that we would strongly benefit from the same approach as in electrostatics, i.e., from breaking Eq. (1) into a product of two factors via the introduction of a suitable *field*. Such decomposition may be done as follows:

$$\mathbf{F} = \int_V \mathbf{j}(\mathbf{r}) \times \mathbf{B}(\mathbf{r}) d^3 r, \quad (5.8)$$

Lorentz
force on
a current

⁵ In particular, Eq. (7) is used for the legal definition of the SI unit of current, one *ampere* (A), via the SI unit of force (the newton, N), with coefficient μ_0 fixed as listed above.

where vector \mathbf{B} is called the *magnetic field* (in our particular case, induced by current \mathbf{j}):⁶

$$\mathbf{B}(\mathbf{r}) \equiv \frac{\mu_0}{4\pi} \int_{V'} \mathbf{j}'(\mathbf{r}') \times \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} d^3 r'. \quad (5.9)$$

Biot-Savart law

The last equation is called the *Biot-Savart law*, while \mathbf{F} expressed by Eq. (8) is sometimes called the *Lorentz force*.⁷ However, more frequently the later term is reserved for the full force,

$$\mathbf{F} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \quad (5.10)$$

Lorentz force on a particle

exerted by electric and magnetic fields on a point charge q , moving with velocity \mathbf{v} . (The equivalence of Eq. (8) and the magnetic part of Eq. (10) follows from the summation of all forces acting on n particles in a unit volume, moving with the same velocity \mathbf{v} , so that $\mathbf{j} = qn\mathbf{v}$.)

Now we have to prove that the new formulation (8)-(9) is equivalent to Eq. (1). At the first glance, this seems unlikely. Indeed, first of all, Eqs. (8) and (9) involve vector products, while Eq. (1) is based on a scalar product. More profoundly, in contrast to Eq. (1), Eqs. (8) and (9) do *not* satisfy the 3rd Newton's law, applied to elementary current components $\mathbf{j} d^3 r$ and $\mathbf{j}' d^3 r'$, if these vectors are not parallel to each other. Indeed, consider the situation shown in Fig. 3. Here vector \mathbf{j}' is perpendicular to vector $(\mathbf{r} - \mathbf{r}')$, and hence, according to Eq. (9), produces a nonvanishing contribution $d\mathbf{B}'$ to the magnetic field, directed (in Fig. 3) perpendicular to the plane of drawing, i.e. is perpendicular to vector \mathbf{j} . Hence, according to Eq. (8), this field provides a nonvanishing contribution to \mathbf{F} . On the other hand, if we calculate the reciprocal force \mathbf{F}' by swapping indices in Eqs. (8) and (9), the latter equation immediately shows that $d\mathbf{B}(\mathbf{r}') \propto \mathbf{j} \times (\mathbf{r} - \mathbf{r}') = 0$, because the two operand vectors are parallel (Fig. 3). Hence, the current component $\mathbf{j}' d^3 r'$ does exert a force on its counterpart, while $\mathbf{j} d^3 r$ does not.

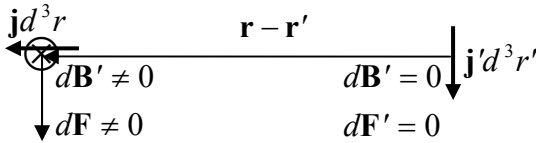


Fig. 5.3. Apparent violation of the 3rd Newton law in magnetism.

Despite this apparent problem, let us still go ahead and plug Eq. (9) into Eq. (8):

$$\mathbf{F} = \frac{\mu_0}{4\pi} \int_V d^3 r \int_{V'} d^3 r' \mathbf{j}(\mathbf{r}) \times \left(\mathbf{j}'(\mathbf{r}') \times \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} \right). \quad (5.11)$$

⁶ The SI unit of the magnetic field is called *tesla*, T - after N. Tesla, an electrical engineering pioneer. In the Gaussian units, the already discussed constant $1/c^2$ in Eq. (1) is equally divided between Eqs. (8) and (9), so that in them both, the constant before the integral is $1/c$. The resulting Gaussian unit of field \mathbf{B} is called *gauss* (G); taking into account the difference of units of electric charge and length, and hence current density, 1 G equals exactly 10^{-4} T. Note also that in some textbooks, especially old ones, \mathbf{B} is called either the *magnetic induction*, or the *magnetic flux density*, while the term “magnetic field” is reserved for vector \mathbf{H} that will be introduced Sec. 5 below.

⁷ Named after H. Lorentz, who received a Nobel prize for his explanation of the Zeeman effect, but is more famous for his numerous contributions to the development of special relativity – see Chapter 9. To be fair, the magnetic part of the Lorentz force was correctly calculated first by O. Heaviside.

This double vector product may be transformed into two scalar products, using the vector algebraic identity called the *bac minus cab rule*, $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{b}(\mathbf{a} \cdot \mathbf{c}) - \mathbf{c}(\mathbf{a} \cdot \mathbf{b})$.⁸ Applying this relation, with $\mathbf{a} = \mathbf{j}$, $\mathbf{b} = \mathbf{j}'$, and $\mathbf{c} = \mathbf{R} \equiv \mathbf{r} - \mathbf{r}'$, to Eq. (11), we get

$$\mathbf{F} = \frac{\mu_0}{4\pi} \int_{V'} d^3 r' \mathbf{j}'(\mathbf{r}') \left(\int_V d^3 r \frac{\mathbf{j}(\mathbf{r}) \cdot \mathbf{R}}{R^3} \right) - \frac{\mu_0}{4\pi} \int_V d^3 r \int_{V'} d^3 r' \mathbf{j}(\mathbf{r}) \cdot \mathbf{j}'(\mathbf{r}') \frac{\mathbf{R}}{R^3}. \quad (5.12)$$

The second term in the right-hand part of this equation coincides with the right-hand part of Eq. (1), while the first term equals zero, because its internal integral vanishes. Indeed, we may break volumes V and V' into narrow *current tubes*, the stretched sub-volumes whose walls are not crossed by current lines ($j_n = 0$). As a result, the (infinitesimal) current in each tube, $dI = j dA = j d^2 r$, is the same along its length, and, just as in a thin wire, $\mathbf{j} d^2 r$ may be replaced with $dI d\mathbf{r}$. Because of this, each tube's contribution to the internal integral in the first term of Eq. (12) may be presented as

$$dI \oint_l d\mathbf{r} \cdot \frac{\mathbf{R}}{R^3} = -dI \oint_l d\mathbf{r} \cdot \nabla \frac{1}{R} = -dI \oint_l d\mathbf{r} \frac{\partial}{\partial r} \frac{1}{R}, \quad (5.13)$$

where operator ∇ acts in the \mathbf{r} space, and the integral is taken along tube's length l . Due to the current continuity, each loop should follow a closed contour, and an integral of a full differential of some scalar function (in our case, $1/r_{12}$) along it equals zero.

So we have recovered Eq. (1). Returning for a minute to the paradox illustrated with Fig. 3, we may conclude that the apparent violation of the 3rd Newton law was the artifact of our interpretation of Eqs. (8) and (9) as sums of independent elementary components. In reality, due to the dc current continuity expressed by Eq. (4.6), these components are *not* independent. For the whole currents, Eqs. (8)-(9) do obey the 3rd law – as follows from their already proved equivalence to Eq. (1).

Thus we have been able to break the magnetic interaction into the two effects: the creation of the *magnetic field* \mathbf{B} by one current (in our notation, \mathbf{j}'), and the effect of this field on the other current (\mathbf{j}). Now comes an additional experimental fact: other elementary components $\mathbf{j} d^3 r'$ of current \mathbf{j} also contribute to the magnetic field (9) acting on component $\mathbf{j} d^3 r$.⁹ This fact allows us to drop prime after \mathbf{j} in Eq. (9), and rewrite Eqs. (8) and (9) as

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int_{V'} \mathbf{j}(\mathbf{r}') \times \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} d^3 r', \quad (5.14)$$

$$\mathbf{F} = \int_V \mathbf{j}(\mathbf{r}) \times \mathbf{B}(\mathbf{r}) d^3 r, \quad (5.15)$$

Again, the field *observation* point \mathbf{r} and the field *source* point \mathbf{r}' have to be clearly distinguished. We immediately see that these expressions are similar to, but still different from the corresponding relations of the electrostatics, namely Eq. (1.8),

⁸ See, e.g., MA Eq. (7.5).

⁹ Just in electrostatics, one needs to exercise due caution at transfer from these expressions to the limit of discrete classical particles, and extended wavefunctions in quantum mechanics, in order to avoid the (non-existing) magnetic interaction of a charged particle upon itself.

$$\mathbf{E}(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \oint_{V'} \rho(\mathbf{r}') \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} d^3r', \quad (5.16)$$

and the distributed version of Eq. (1.6):

$$\mathbf{F} = \oint_V \rho(\mathbf{r}) \mathbf{E}(\mathbf{r}) d^3r. \quad (5.17)$$

(Note that the sign difference has disappeared, at the cost of the replacement of scalar-by-vector multiplications in electrostatics with cross-products of vectors in magnetostatics.)

For the frequent case of a field of a thin wire of length l' , Eq. (14) may be re-written as

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0 I}{4\pi} \oint_{l'} d\mathbf{r}' \times \frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3}. \quad (5.18)$$

Let us see how does the last formula work for the simplest case of a straight wire (Fig. 4a). The magnetic field contribution $d\mathbf{B}$ due to any small fragment $d\mathbf{r}'$ of the wire's length is directed along the same line (perpendicular to both the wire and the perpendicular d dropped from the observation point to the wire line), and its magnitude is

$$dB = \frac{\mu_0 I}{4\pi} \frac{dx'}{|\mathbf{r} - \mathbf{r}'|^2} \sin \theta = \frac{\mu_0 I}{4\pi} \frac{dx'}{(d^2 + x^2)} \frac{d}{(d^2 + x^2)^{1/2}}. \quad (5.19)$$

Summing up all such contributions, we get

$$B = \frac{\mu_0 I \rho}{4\pi} \int_{-\infty}^{\infty} \frac{dx}{(x^2 + d^2)^{3/2}} = \frac{\mu_0 I}{2\pi d}. \quad (5.20)$$

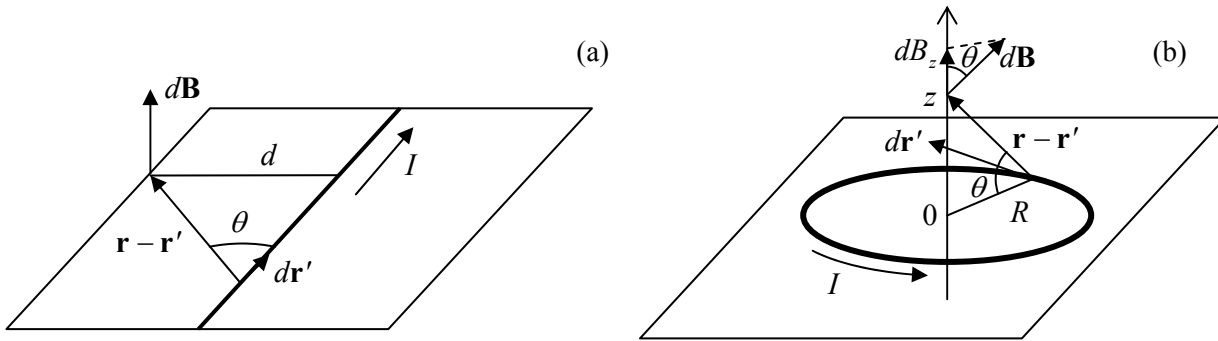


Fig. 5.4. Magnetic fields of: (a) a straight current, and (b) a current loop.

This is a simple but very important result. (Note that it is only valid for very long ($l \gg d$), straight wires.) It is especially crucial to note the “vortex” character of the field: its lines go around the wire, forming round rings with the centers on the current line. This is in the sharp contrast to the electrostatic field lines that can only begin and end on electric charges and never form closed loops (otherwise the Coulomb force $q\mathbf{E}$ would not be conservative). In the magnetic case, the vortex *field* may be reconciled with the potential character of magnetic *forces*, which is evident from Eq. (1), due to the vector products in Eqs. (14)-(15).

Now we may use Eq. (15), or rather its thin-wire version

$$\mathbf{F} = I \oint_l d\mathbf{r} \times \mathbf{B}(\mathbf{r}), \quad (5.21)$$

to apply Eq. (20) to the two-wire problem (Fig. 2). Since for the second wire vectors $d\mathbf{r}$ and \mathbf{B} are perpendicular to each other, we immediately arrive at our previous result (7).

The next important application of the Biot-Savart law (14) is the magnetic field at the axis of a circular current loop (Fig. 4b). Due to the problem symmetry, the net field \mathbf{B} has to be directed along the axis, but each of its components $d\mathbf{B}$ is tilted by angle $\theta = \arctan(z/R)$ to this axis, so that its axial component

$$dB_z = dB \cos \theta = \frac{\mu_0 I}{4\pi} \frac{dr'}{R^2 + z^2} \frac{R}{(R^2 + z^2)^{1/2}}. \quad (5.22)$$

Since the denominator of this expression remains the same for all wire components dr' , in this case the integration is trivial ($\oint dr' = 2\pi R$), giving finally

$$B = \frac{\mu_0 I}{2} \frac{R^2}{(R^2 + z^2)^{3/2}}. \quad (5.23)$$

Note that the magnetic field in the loop's center (i.e., for $z = 0$),

$$B = \frac{\mu_0 I}{2R}, \quad (5.24)$$

is π times higher than that due to a similar current in a straight wire, at distance $d = R$ from it. This increase is readily understandable, since all elementary components of the loop are at the same distance R from the observation point, while in the case of a straight wire, all its point but one are separated from the observation point by a distance larger than d .

Another notable fact is that at large distances ($z^2 \gg R^2$), field (23) is proportional to z^{-3} :

$$B \approx \frac{\mu_0 I}{2} \frac{R^2}{|z|^3} = \frac{\mu_0}{4\pi} \frac{2m}{|z|^3}, \quad (5.25)$$

just like the electric field of a dipole (along its direction), with the replacement of the electric dipole moment magnitude p with $m = IA$, where $A = \pi R^2$ is the loop area. This is the best example of a *magnetic dipole*, with *dipole moment* m - the notions to be discussed in more detail in Sec. 5 below.

5.2. Vector-potential and the Ampère law

The reader can see that the calculations of the magnetic field using Eq. (14) or (18) are still cumbersome even for the very simple systems we have examined. As we saw in Chapter 1, similar calculations in electrostatics, at least for several important systems of high symmetry, could be substantially simplified using the Gauss law (1.16). A similar relation exists in magnetostatics as well, but has a different form, due to the vortex character of the magnetic field. To derive it, let us notice that in an analogy with the scalar case, the vector product under integral (14) may be transformed as

$$\frac{\mathbf{j}(\mathbf{r}') \times (\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} = \nabla \times \frac{\mathbf{j}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}, \quad (5.26)$$

where operator ∇ acts in the \mathbf{r} space. (This equality may be really verified by its Cartesian components, noticing that the current density is a function of \mathbf{r}' and hence its components are independent of \mathbf{r} .) Plugging Eq. (26) into Eq. (14), and moving operator ∇ out of the integral over \mathbf{r}' , we see that the magnetic field may be presented as the curl of another vector field:¹⁰

$$\mathbf{B}(\mathbf{r}) = \nabla \times \mathbf{A}(\mathbf{r}), \quad (5.27)$$

namely the so-called *vector-potential*:

$$\mathbf{A}(\mathbf{r}) \equiv \frac{\mu_0}{4\pi} \int_{V'} \frac{\mathbf{j}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3 r'. \quad (5.28)$$

Vector-
potential

Please note a wonderful analogy between Eqs. (27)-(28) and, respectively, Eqs. (1.33) and (1.38). This analogy implies that vector-potential \mathbf{A} plays, for the magnetic field, essentially the same role as the scalar potential ϕ plays for the electric field (hence the name “potential”), with due respect to the vortex character of \mathbf{A} . I will discuss this notion in detail below.

Now let us see what equations we may get for the spatial derivatives of the magnetic field. First, vector algebra says that the divergence of any curl is zero.¹¹ In application to Eq. (27), this means that

$$\nabla \cdot \mathbf{B} = 0. \quad (5.29)$$

No
magnetic
monopoles

Comparing this equation with Eq. (1.27), we see that Eq. (29) may be interpreted as the absence of a magnetic analog of an electric charge on which magnetic field lines could originate or end. Numerous searches for such hypothetical magnetic charges, called *magnetic monopoles*, using very sensitive and sophisticated experimental setups, have never given a convincing evidence of their existence in Nature.

Proceeding to the alternative, vector derivative of the magnetic field (i.e., its curl), and using Eq. (28), we get

$$\nabla \times \mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \nabla \times \left(\nabla \times \int_{V'} \frac{\mathbf{j}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3 r' \right). \quad (5.30)$$

This expression may be simplified by using the following general vector identity:¹²

$$\nabla \times (\nabla \times \mathbf{c}) = \nabla(\nabla \cdot \mathbf{c}) - \nabla^2 \mathbf{c}, \quad (5.31)$$

applied to vector $\mathbf{c}(\mathbf{r}) = \mathbf{j}(\mathbf{r}')/|\mathbf{r} - \mathbf{r}'|$:

$$\nabla \times \mathbf{B} = \frac{\mu_0}{4\pi} \nabla \int_{V'} \mathbf{j}(\mathbf{r}') \cdot \nabla \frac{1}{|\mathbf{r} - \mathbf{r}'|} d^3 r' - \frac{\mu_0}{4\pi} \int_{V'} \mathbf{j}(\mathbf{r}') \nabla^2 \frac{1}{|\mathbf{r} - \mathbf{r}'|} d^3 r'. \quad (5.32)$$

As was already discussed during our study of electrostatics,

¹⁰ In the Gaussian units, Eq. (27) remains the same, and hence in Eq. (28), coefficient $\mu_0/4\pi$ is replaced with $1/c$.

¹¹ See, e.g., MA Eq. (11.2).

¹² See, e.g., MA Eq. (11.3).

$$\nabla^2 \frac{1}{|\mathbf{r} - \mathbf{r}'|} = -4\pi\delta(\mathbf{r} - \mathbf{r}'), \quad (5.33)$$

so that the last term of Eq. (32) is just $\mu_0 \mathbf{j}(\mathbf{r})$. On the other hand, inside the first integral we can replace ∇ with $(-\nabla')$, where prime means differentiation in the space of radius-vector \mathbf{r}' . Integrating that term by parts, we get

$$\nabla \times \mathbf{B} = -\frac{\mu_0}{4\pi} \nabla \oint_{S'} j_n(\mathbf{r}') \frac{1}{|\mathbf{r} - \mathbf{r}'|} d^2 r' + \nabla \int_{V'} \frac{\nabla' \cdot \mathbf{j}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3 r' + \mu_0 \mathbf{j}(\mathbf{r}). \quad (5.34)$$

Applying this equation to the volume V' limited by a surface S' sufficiently distant from the field concentration (or with no current crossing it), we may neglect the first term in the right-hand part of Eq. (34), while the second term always equals zero in statics, due to the dc charge continuity – see Eq. (4.6). As a result, we arrive at a very simple differential equation¹³

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j}. \quad (5.35)$$

This is (the dc form of) the inhomogeneous Maxwell equation, which in magnetostatics plays the role similar to the Poisson equation (1.27) in electrostatics. Let me display, for the first time in this course, this fundamental system of equations (at this stage, for statics only), and give the reader a minute to stare at their beautiful symmetry - that has inspired so much of the 20th century physics:

$$\begin{aligned} \nabla \times \mathbf{E} &= 0, & \nabla \times \mathbf{B} &= \mu_0 \mathbf{j}, \\ \nabla \cdot \mathbf{E} &= \frac{\rho}{\varepsilon_0}, & \nabla \cdot \mathbf{B} &= 0. \end{aligned} \quad (5.36)$$

Static
Maxwell
equations

Their only asymmetry, two zeros in the right hand parts (for the magnetic field's divergence and electric field's curl), is due to the absence in Nature of, respectively, the magnetic monopoles and their currents. I will discuss these equations in more detail in Sec. 6.7, after the equations for field curls have been generalized to their full (time-dependent) versions.

Returning now to a more mundane but important task of calculating magnetic field induced by simple current configurations, we can benefit from an integral form of Eq. (35). For that, let us integrate this equation over an arbitrary surface S limited by a closed contour C , applying to it the Stokes theorem.¹⁴ The resulting expression,

$$\oint_C \mathbf{B} \cdot d\mathbf{r} = \mu_0 \oint_S j_n d^2 r \equiv \mu_0 I, \quad (5.37)$$

Ampère
law

where I is the net electric current crossing surface S , is called the *Ampère law*.

As the first example of its application, let us return to a current in a straight wire (Fig. 4a). With the Ampère law in our arsenal, we can readily pursue an even more ambitious goal – calculate the magnetic field both outside and inside of a wire of arbitrary radius R , with an arbitrary (albeit axially-symmetric) current distribution $j(\rho)$ – see Fig. 5. Selecting two contours C in the form of rings of some

¹³ As in all earlier formulas for the magnetic field, in the Gaussian units the coefficient μ_0 in this relation has to be replaced with $4\pi/c$.

¹⁴ See, e.g., MA Eq. (12.1) with $\mathbf{f} = \mathbf{B}$.

radius ρ in the plane perpendicular to the wire axis z , we have $\mathbf{B} \cdot d\mathbf{r} = B\rho(d\varphi)$, these φ is the azimuthal angle, so that the Ampère law (37) yields:

$$2\pi \rho B = \mu_0 \times \begin{cases} 2\pi \int_0^\rho j(\rho') \rho' d\rho', & \text{for } \rho \leq R, \\ 2\pi \int_0^R j(\rho') \rho' d\rho' \equiv I, & \text{for } \rho \geq R. \end{cases} \quad (5.38)$$

Thus we have not only recovered our previous result (20), with the notation replacement $d \rightarrow \rho$, in a much simpler way, but could also find the magnetic field distribution inside the wire. (In the most common case when the wire conductivity σ is constant, and hence the current is uniformly distributed along its cross-section, $j(\rho) = \text{const}$, the first of Eqs. (38) immediately yields $B \propto \rho$ for $\rho \leq R$).

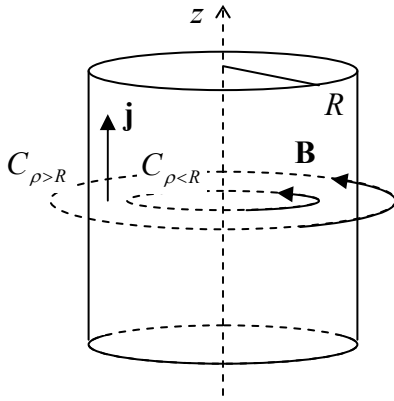


Fig. 5.5. The simplest application of the Ampère law: dc current in a straight wire.

Another important example is a straight, long *solenoid* (Fig. 6a), with dense winding: $n^2 A \gg 1$, where n is the number of wire turns per unit length and A is the area of solenoid's cross-section - not necessarily circular.

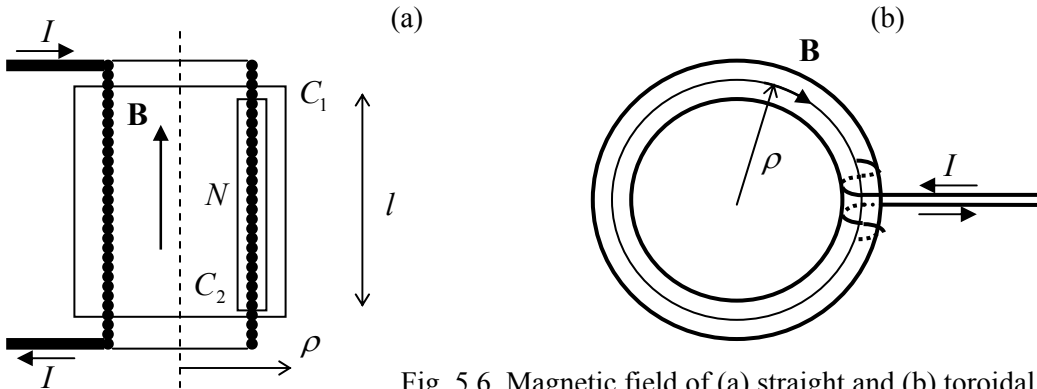


Fig. 5.6. Magnetic field of (a) straight and (b) toroidal solenoids.

From the symmetry of this problem, the longitudinal (in Fig. 6a, vertical) component B_z of the magnetic field may only depend on the horizontal position \mathbf{p} of the observation point. First taking a plane Ampère contour C_1 , with both long sides outside the solenoid, we get $B_z(\mathbf{p}_2) - B_z(\mathbf{p}_1) = 0$, because the total current piercing the contour equals zero. This is only possible if $B_z = 0$ at any ρ outside of the

(infinitely long!) solenoid.¹⁵ With this result on hand, from contour C_2 we get the following relation for the only (z -) component of the internal field:

$$Bl = \mu_0 NI, \quad (5.39)$$

where N is the number of wire turns passing through the contour of length l . This means that regardless of the exact position internal side of the contour, the result is the same:

$$B = \mu_0 \frac{N}{l} I = \mu_0 nI. \quad (5.40)$$

Thus, the field inside an infinitely long solenoid is uniform; in this sense, a long solenoid is a magnetic analog of a wide plane capacitor.

As should be clear from its derivation, the obtained result, especially that the field outside of the solenoid equals zero, is conditional on the solenoid length being very large in comparison with its lateral size. (From Eq. (25), we may predict that for a solenoid of a finite length l , the external field is only a factor of $\sim A/l^2$ lower than the internal one.) Much better suppression of this external (“fringe”) field may be obtained using the *toroidal solenoid* (Fig. 6b). The application of Ampère law to this geometry shows that, in the limit of dense winding ($N \gg 1$), there is no fringe field at all – for any relation between two radii of the torus, while inside the solenoid, and distance ρ from the center,

$$B = \frac{\mu_0 NI}{2\pi\rho}. \quad (5.41)$$

We see that a possible drawback of this system for practical applications is that internal field depends on ρ , i.e. is not quite uniform; however, if the torus is thin, this problem is minor.

How should we solve the problems of magnetostatics for systems whose low symmetry does not allow getting easy results from the Ampère law? (The examples are of course too numerous to list; for example, we cannot use this approach even to reproduce Eq. (23) for a round current loop.) From the deep analogy with electrostatics, we may expect that in this case we could recover the field from the solution of a certain partial boundary problem for the field’s potential, in this case the vector-potential \mathbf{A} defined by Eq. (28). However, despite the similarity of this formula and Eq. (1.38) for ϕ , that was emphasized above, there are two additional issues we should tackle in the magnetic case.

First, finding vector-potential distribution means determining three scalar functions (say, A_x , A_y , and A_z), rather than one (ϕ). Second, generally the differential equation satisfied by \mathbf{A} is more complex than the Poisson equation for ϕ . Indeed, plugging Eq. (27) into Eq. (35), we get

$$\nabla \times (\nabla \times \mathbf{A}) = \mu_0 \mathbf{j}. \quad (5.42)$$

If we wrote the left-hand part of this equation in (say, Cartesian) components, we would see that they are much more interwoven than in the Laplace operator, and hence much less convenient for using the orthogonal coordinate approach or the variable separation method. In order to remedy the situation, let us apply to Eq. (42) the now-familiar identity (31). The result is

¹⁵ Applying the Ampère law to a circular contour of radius ρ , coaxial with the solenoid, we see that the field outside (but not inside!) it has an azimuthal component B_ϕ , similar to that of the straight wire (see Eq. (38) above) and hence (at $N \gg 1$) much weaker than the longitudinal field inside the solenoid – see Eq. (40).

$$\nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} = \mu_0 \mathbf{j}. \quad (5.43)$$

We see that if we could kill the first term in the left-hand part, for example if $\nabla \cdot \mathbf{A} = 0$, the second term would give us a set of independent Poisson equations for each Cartesian component of vector \mathbf{A} .

In this context, let us discuss what discretion do we have in the potential choice. In electrostatics, we might add to ϕ not only an arbitrary constant, but also an arbitrary function of time, without affecting the electric field:

$$-\nabla[\phi + f(t)] = -\nabla\phi = \mathbf{E}. \quad (5.44)$$

Similarly, using the fact that curl of the gradient of any scalar function equals zero,¹⁶ we may add to \mathbf{A} not only a constant, but even a gradient of an arbitrary function $\chi(\mathbf{r}, t)$, because

$$\nabla \times (\mathbf{A} + \nabla\chi) = \nabla \times \mathbf{A} + \nabla \times (\nabla\chi) = \nabla \times \mathbf{A} = \mathbf{B}. \quad (5.45)$$

Such additions, keeping the actual (observable) fields intact, are called *gauge transformations*.¹⁷ Let us see what such a transformation does to $\nabla \cdot \mathbf{A}$:

$$\nabla \cdot (\mathbf{A} + \nabla\chi) = \nabla \cdot \mathbf{A} + \nabla^2 \chi. \quad (5.46)$$

Hence we can choose a function χ in such a way that the divergence of the transformed vector-potential, $\mathbf{A}' \equiv \mathbf{A} + \nabla\chi$, would vanish, so that the new vector-potential would satisfy the vector Poisson equation

$$\nabla^2 \mathbf{A}' = -\mu_0 \mathbf{j}, \quad (5.47)$$

Poisson
equation
for \mathbf{A}

together with the so-called *Coulomb gauge* condition:

$$\nabla \cdot \mathbf{A}' = 0. \quad (5.48)$$

Coulomb
gauge

This gauge is very convenient; one should, however, remember that the resulting solution $\mathbf{A}'(\mathbf{r})$ may differ from the function given by Eq. (28) - while field \mathbf{B} remains the same.¹⁸

In order to get a better feeling of vector-potential's distribution in space, let us solve Eq. (47) for the same straight wire problem (Fig. 5). As Eq. (28) shows, in this case vector \mathbf{A} has just one component (along the axis z). Moreover, due to the problem's axial symmetry, its magnitude may only depend on the distance from the axis: $\mathbf{A} = \mathbf{n}_z A(\rho)$. Hence, the gradient of \mathbf{A} is directed across axis z , so that Eq. (48) is satisfied even for this vector, i.e. the Poisson equation (47) is satisfied even for the original vector \mathbf{A} . For our symmetry ($\partial/\partial\phi = \partial/\partial z = 0$), the Laplace operator, written in cylindrical coordinates, has just one term,¹⁹ reducing Eq. (47) to

$$\frac{1}{\rho} \frac{d}{d\rho} \left(\rho \frac{dA}{d\rho} \right) = -\mu_0 j(\rho). \quad (5.49)$$

Multiplying both parts of this equation by ρ and integrating them over the coordinate once, we get

¹⁶ See, e.g., MA Eq. (11.1).

¹⁷ The use of term “gauge” (originally meaning “a measure” or “a scale”) in this context is purely historic, so the reader should not try to find too much hidden sense in it.

¹⁸ Since most equations for \mathbf{A} are valid for \mathbf{A}' as well, I will follow the common (possibly, bad) tradition, and in many cases use the same notation, \mathbf{A} , for both functions.

¹⁹ See, e.g., MA Eq. (10.3).

$$\rho \frac{dA}{d\rho} = -\mu_0 \int_0^\rho j(\rho') \rho' d\rho' + \text{const.} \quad (5.50)$$

Since in the cylindrical coordinates, for our symmetry,²⁰ $B = -dA/d\rho$, Eq. (50) is nothing else than our old result (38) for the magnetic field.²¹ However, let us continue the integration, at least for the region outside the wire, where the function $A(\rho)$ depends only on the full current I rather than on the current distribution inside the wire. Dividing both parts of Eq. (50) by ρ , and integrating them over that coordinate again, we get

$$A(\rho) = -\frac{\mu_0 I}{2\pi} \ln \rho + \text{const}, \quad \text{where } I = 2\pi \int_0^R j(\rho) \rho d\rho. \quad (5.51)$$

As a reminder, we had the similar logarithmic behavior for the electrostatic potential outside a uniformly charged straight line. This is natural, because the Poisson equations for both cases are similar.

Now let us find the vector-potential for the long solenoid (Fig. 6a), with its uniform magnetic field. Since Eq. (28) prescribes vector \mathbf{A} to follow the direction of the current, we can start with looking for it in the form $\mathbf{A} = \mathbf{n}_\varphi A(\rho)$. (This is especially natural if the solenoid's cross-section is circular.) With this orientation of \mathbf{A} , the same general expression for the curl operator in cylindrical coordinates yields $\nabla \times \mathbf{A} = \mathbf{n}_z (1/\rho) d(\rho A)/d\rho$. According to the definition (27) of \mathbf{A} , this expression should be equal to \mathbf{B} , in our case equal to $\mathbf{n}_z B$, with constant B – see Eq. (40). Integrating this equality, and selecting such integration constant so that $A(0)$ is finite, we get

$$A(\rho) = \frac{B\rho}{2}. \quad (5.52)$$

Plugging this result into the general expression for the Laplace operator in the cylindrical coordinates,²² we see that the Poisson equation (47) with $\mathbf{j} = 0$ (i.e. the Laplace equation), is satisfied again – which is natural since for this distribution, $\nabla \cdot \mathbf{A} = 0$. However, Eq. (52) is not the unique (or even the simplest) solution of the problem. Indeed, using the well-known expression for the curl operator in Cartesian coordinates,²³ it is straightforward to check that either function $\mathbf{A}' = \mathbf{n}_y Bx$, or function $\mathbf{A}'' = -\mathbf{n}_x By$, or any of their weighed sums, for example $\mathbf{A}''' = (\mathbf{A}' + \mathbf{A}'')/2 = B(-\mathbf{n}_x y + \mathbf{n}_y x)/2$, also give the same magnetic field, and also evidently satisfy the Laplace equation. If such solutions do not look very natural due to their anisotropy in the $[x, y]$ plane, please consider the fact that they represent the uniform magnetic field regardless of its source (e.g., of the shape of long solenoid's cross-section). Such choices of vector-potential may be very convenient for some problems, for example for the analysis of the 2D motion of a charged quantum particle in the perpendicular magnetic field, giving the famous Landau energy levels.²⁴

²⁰ See, e.g., MA Eq. (10.5) with $\partial/\partial\varphi = \partial/\partial z = 0$.

²¹ Since the magnetic field at the wire axis has to be zero (otherwise, being perpendicular to the axis, where would it be directed?), the integration constant in Eq. (50) should be zero.

²² See, e.g., MA Eq. (10.6).

²³ See, e.g., MA Eq. (8.5).

²⁴ See, e.g., QM Sec. 3.2.

5.3. Magnetic energy, flux, and inductance

Considering currents flowing in a system as generalized coordinates, magnetic forces (1) between them are their unique functions, and in this sense the magnetic interaction energy U may be considered a potential energy of the system. The apparent (but deceptive) way to guess the energy is to use the analogy between Eq. (1) and its electrostatic analog, Eq. (2). As we know from Chapter 1, if these densities describe the distribution of the same charge, i.e. if $\rho'(\mathbf{r}) = \rho(\mathbf{r})$, then the self-interaction of its elementary components correspond to the potential energy expressed by Eq. (1.61):

$$U = \frac{1}{4\pi\epsilon_0} \frac{1}{2} \int d^3r \int d^3r' \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}. \quad (5.53)$$

Using the analogy, for the magnetic interaction between elementary components of the same current, with density $\mathbf{j}(\mathbf{r}) = \mathbf{j}'(\mathbf{r})$, we could guess that

$$U = \frac{\mu_0}{4\pi} \frac{1}{2} \int d^3r \int d^3r' \frac{\mathbf{j}(\mathbf{r}) \cdot \mathbf{j}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}, \quad (5.53)$$

Magnetic
interaction
energy

while for independent currents the coefficient $\frac{1}{2}$ should be removed. Now let me confess that this is a *wrong* way to get this *correct* result. Indeed, the sign in Eq. (1) is opposite to that in Eq. (2), so that following this argumentation we would get Eq. (53) with the minus sign. The reason of this paradox is fundamental: fixing electric charges does not require external interference (work), while the maintenance of currents generally does. Strictly speaking, a derivation of Eq. (53) required additional experimental fact, the *Faraday induction law*. However, I would like to defer its discussion until the beginning of the next chapter, and for now ask the reader to believe me that the sign in Eq. (53) is correct.

Due to the importance of this relation, let us rewrite it in several other forms, beneficial for different applications. First of all, just as in electrostatics, Eq. (54) may be recast into a potential-based form. Indeed, using definition (28) of the vector-potential $\mathbf{A}(\mathbf{r})$, Eq. (54) becomes²⁵

$$U = \frac{1}{2} \int \mathbf{j}(\mathbf{r}) \cdot \mathbf{A}(\mathbf{r}) d^3r. \quad (5.55)$$

This formula, that is a clear magnetic analog of Eq. (1.62) of electrostatics, is very popular among theoretical physicists, because it is very handy for the field theory manipulations. However, for many calculations it is more convenient to have a direct expression of energy via the magnetic field. Again, this may be done very similarly to what we have done in Sec. 1.3 for electrostatics, i.e. plugging into Eq. (55) the current density expressed from Eq. (35) to transform it as²⁶

$$U = \frac{1}{2} \int \mathbf{j} \cdot \mathbf{A} d^3r = \frac{1}{2\mu_0} \int \mathbf{A} \cdot (\nabla \times \mathbf{B}) d^3r = \frac{1}{2\mu_0} \int \mathbf{B} \cdot (\nabla \times \mathbf{A}) d^3r - \frac{1}{2\mu_0} \int \nabla \cdot (\mathbf{A} \times \mathbf{B}) d^3r. \quad (5.56)$$

Now using the divergence theorem, the second integral may be transformed into a surface integral of product $(\mathbf{A} \times \mathbf{B})_n$. Equations (27)-(28) show that if the current distribution $\mathbf{j}(\mathbf{r})$ is localized, this product drops with distance r faster than $1/r^2$, so that if the integration volume is large enough, the surface

²⁵ This relation remains the same in the Gaussian units, because in those units both Eq. (28) and Eq. (54) should be stripped of their $\mu_0/4\pi$ coefficients.

²⁶ For that, we may use MA Eq. (11.7) with $\mathbf{f} = \mathbf{A}$ and $\mathbf{g} = \mathbf{B}$, giving $\mathbf{A} \cdot (\nabla \times \mathbf{B}) = \mathbf{B} \cdot (\nabla \times \mathbf{A}) - \nabla \cdot (\mathbf{A} \times \mathbf{B})$.

integral is negligible. In the remaining first integral, we may use Eq. (27) to recast $\nabla \times \mathbf{A}$ into the magnetic field. As a result, we get a very simple and fundamental formula.

$$U = \frac{1}{2\mu_0} \int B^2 d^3r. \quad (5.57a)$$

Just as with the electric field, this expression may be interpreted as a volume integral of the *magnetic energy density* u :

Magnetic
field
energy

$$U = \int u(\mathbf{r}) d^3r, \quad \text{with } u(\mathbf{r}) \equiv \frac{1}{2\mu_0} \mathbf{B}^2(\mathbf{r}), \quad (5.57b)$$

clearly similar to Eq. (1.67).²⁷ Again, the conceptual choice between the spatial localization of magnetic energy – either at the location of electric currents only, as implied by Eqs. (54) and (55), or in all regions where the magnetic field exists, as apparent from Eq. (57b), cannot be done within the framework of magnetostatics, and only electrodynamics gives the decisive preference for the latter choice.

For the practically important case of currents flowing in several thin wires, Eq. (54) may be first integrated over the cross-section of each wire, just as was done at the derivation of Eq. (4). Again, since the integral of the current density over k^{th} wire's cross-section is just the current I_k in the wire, and cannot change along its length, it may be taken from the remaining integrals, giving

$$U = \frac{\mu_0}{4\pi} \frac{1}{2} \sum_{k,k'} I_k I_{k'} \oint_{l_k} \oint_{l_{k'}} \frac{d\mathbf{r}_k \cdot d\mathbf{r}_{k'}}{|\mathbf{r}_k - \mathbf{r}_{k'}|}, \quad (5.58)$$

where l is the full length of the wire loop. Note that Eq. (58) is valid if currents I_k are independent of each other, because the double sum counts each current pair twice, compensating coefficient $\frac{1}{2}$ in front of the sum. It is useful to decompose this relation as

$$U = \frac{1}{2} \sum_{k,k'} I_k I_{k'} L_{kk'}, \quad (5.59)$$

Mutual
inductance
coefficients

$$L_{kk'} \equiv \frac{\mu_0}{4\pi} \oint_{l_k} \oint_{l_{k'}} \frac{d\mathbf{r}_k \cdot d\mathbf{r}_{k'}}{|\mathbf{r}_k - \mathbf{r}_{k'}|}, \quad (5.60)$$

Coefficient $L_{kk'}$ in the quadratic form (59), with $k \neq k'$, is called the *mutual inductance* between current loops k and k' , while the diagonal coefficient $L_k \equiv L_{kk}$ is called the *self-inductance* (or just *inductance*) of k^{th} loop.²⁸ From the symmetry of Eq. (60) with respect to the index swap, $k \leftrightarrow k'$, it evident that the matrix of coefficients $L_{kk'}$ is symmetric.²⁹

²⁷ The transfer to the Gaussian units in Eqs. (77)-(78) may be accomplished by the usual replacement $\mu_0 \rightarrow 4\pi$, thus giving, in particular, $u = B^2/8\pi$.

²⁸ As evident from Eq. (60), these coefficients depend only on the geometry of the system. Moreover, in the Gaussian units, in which Eq. (60) is valid without the factor $\mu_0/4\pi$, the inductance coefficients have the dimension of length (centimeters). The SI unit of inductance is called the *henry*, abbreviated H - after J. Henry, 1797-1878, who in particular discovered the effect of electromagnetic induction (see Sec. 6.1) independently of M. Faraday.

²⁹ Note that the matrix of the mutual inductances $L_{jj'}$ is very much similar to the matrix of *reciprocal* capacitance coefficients $p_{kk'}$ – for example, compare Eq. (62) with Eq. (2.21).

$$L_{kk'} = L_{k'k}, \quad (5.61)$$

so that for the practically important case of two interacting currents I_1 and I_2 , Eq. (59) reads

$$U = \frac{1}{2} L_1 I_1^2 + M I_1 I_2 + \frac{1}{2} L_2 I_2^2, \quad (5.62)$$

where $M \equiv L_{12} = L_{21}$ is the mutual inductance coefficient.

These formulas clearly show the importance of self- and mutual inductances, so I will demonstrate their calculation for at least a few basic geometries. Before doing that, however, let me recast Eq. (58) into one more form that may facilitate such calculations. Namely, let us notice that for the magnetic field induced by current I_k in a thin wire, Eq. (28) is reduced to

$$\mathbf{A}_k(\mathbf{r}) = \frac{\mu_0}{4\pi} I_k \int_{l'} \frac{d\mathbf{r}_k}{|\mathbf{r} - \mathbf{r}_k|}, \quad (5.63)$$

so that Eq. (58) may be rewritten as

$$U = \frac{1}{2} \sum_{k,k'} I_k \oint_{l_k} \mathbf{A}_{k'}(\mathbf{r}_k) \cdot d\mathbf{r}_{k'}. \quad (5.64)$$

But according to the same Stokes theorem that was used earlier in this chapter to derive the Ampère law, and Eq. (27), such integral is nothing more than the *magnetic field flux* (more frequently called just the *magnetic flux*) through a surface S limited by the contour l :³⁰

$$\oint_l \mathbf{A}(\mathbf{r}) \cdot d\mathbf{r} = \int_S (\nabla \times \mathbf{A})_n d^2r = \int_S B_n d^2r \equiv \Phi. \quad (5.65)$$

Magnetic flux

As a result, Eq. (64) may be rewritten as

$$U = \frac{1}{2} \sum_{k,k'} I_k \Phi_{kk'}, \quad (5.66)$$

where $\Phi_{kk'}$ is the flux of the field induced by k' -th current through the loop of the k -th current. Comparing this expression with Eq. (59), we see that

$$\Phi_{kk'} \equiv \int_{S_k} (\mathbf{B}_{k'})_n d^2r = L_{kk'} I_{k'}, \quad (5.67)$$

Magnetic flux from currents

This expression not only gives us one more means for calculating coefficients $L_{kk'}$, but also shows their physical sense: the mutual inductance characterizes how much field (colloquially, “how many field lines”) induced by current $I_{k'}$ penetrate the loop of current I_k , and vice versa. Since due to the linear superposition principle, the total flux piercing k -th loop may be presented as

$$\Phi_k \equiv \sum_{k'} \Phi_{kk'} = \sum_{k'} L_{kk'} I_{k'}. \quad (5.68)$$

³⁰ The SI unit of magnetic flux is called *weber*, abbreviated Wb - after W. Weber, who in particular co-invented (with C. Gauss) the electromagnetic telegraph, and in 1856 was first, together with R. Kohlrausch, to notice that the value of (in modern terms) $1/(\epsilon_0 \mu_0)^{1/2}$, derived from electrostatic and magnetostatic measurements, coincides with the independently measured speed of light c , giving an important motivation for Maxwell's theory.

For example, for the system of two currents this expression is reduced to a clear analog of Eqs. (2.19):

$$\begin{aligned}\Phi_1 &= L_1 I_1 + M I_2, \\ \Phi_2 &= M I_1 + L_2 I_2.\end{aligned}\quad (5.69)$$

For the even simpler case of a single current,

$$\Phi = L I, \quad (5.70)$$

Φ and U
of a
single
current

so that the magnetic energy of the current may be presented in several equivalent forms:

$$U = \frac{L}{2} I^2 = \frac{1}{2} I \Phi = \frac{1}{2L} \Phi^2. \quad (5.71)$$

These relations, similar to Eqs. (2.14)-(2.15) of electrostatics, show that the self-inductance L of a current loop may be considered as a measure of system's magnetic energy at fixed current.

Now we are well equipped for the calculation of inductances, having three options. The first one is to use Eq. (60) directly.³¹ The second one is to calculate the magnetic field energy from Eq. (57) as the function of currents I_k in the system, and then use Eq. (59) to find all coefficients L_{kk} . For example, for a system with just one current, Eq. (71) yields

$$L = \frac{U}{I^2/2}. \quad (5.72)$$

Finally, if the system consists of thin wires, so that the loop areas S_k and hence fluxes Φ_{kk} are well defined, we may calculate them from Eq. (65), and then use Eq. (67) to find the inductances.

Actually, the first two options may have advantages over the third one even for such system of thin wires for whom the notion of magnetic flux is not quite clear. As an important example, let us find inductance of a long solenoid - see Fig. 6a. We have already calculated the magnetic field inside it - see Eq. (40) - so that, due to the field uniformity, the magnetic flux piercing each wire turn is just

$$\Phi_1 = BA = \mu_0 n I A, \quad (5.73)$$

where A is the area of solenoid's cross-section - for example πR^2 for a round solenoid, though Eq. (40) is more general. Comparing Eqs. (73) and (67), one might wrongly conclude that $L = \Phi_1/I = \mu_0 n A$ **[WRONG!]**, i.e. that the solenoid's inductance is independent on its length. Actually, the magnetic flux Φ_1 pierces *each* wire turn, so that the total flux through the *whole* current loop, consisting of N turns, is

$$\Phi = N \Phi_1 = \mu_0 n^2 l A I, \quad (5.74)$$

and the correct expression for solenoid's inductance is

$$L = \frac{\Phi}{I} = \mu_0 n^2 l A, \quad (5.75)$$

i.e. the inductance per unit length is constant: $L/l = \mu_0 n^2 A$. Since this reasoning may seem a bit flimsy, it is prudent to verify it by using Eq. (72) to calculate the full magnetic energy inside the solenoid (neglecting minor fringe and external field contributions):

³¹ Numerous applications of this *Neumann formula* to electrical engineering problems may be found, for example, in the classical text F. Grover, *Inductance Calculations*, Dover, 1946.

$$U = \frac{1}{2\mu_0} B^2 Al = \frac{1}{2\mu_0} (\mu_0 nI)^2 Al = \mu_0 n^2 l A \frac{I^2}{2}. \quad (5.76)$$

Plugging this result into Eq. (72) immediately confirms result (75).

The use of the first two options for inductance calculation becomes inevitable for continuously distributed currents. As an example, let us calculate self-inductance L of a long coaxial cable with the cross-section shown in the Fig. 7.³²

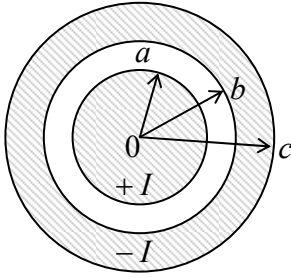


Fig. 5.7. Cross-section of a coaxial cable.

Let us assume that the current is uniformly distributed over the cross-sections of both conductors. (As we know from the previous chapter, such distribution indeed takes place if both the internal and external conductors are made of a uniform resistive material.) First, we should calculate the radial distribution of the magnetic field (that of course has only one, azimuthal component, because of the axial symmetry of the problem). This distribution may be immediately found from the application of the Ampère law to circles of radii ρ within four different ranges:

$$2\pi\rho B = \mu_0 I \Big|_{\text{piercing the circle area}} = \mu_0 I \times \begin{cases} \frac{\rho^2}{a^2}, & \text{for } \rho < a, \\ 1, & \text{for } a < \rho < b, \\ \frac{c^2 - \rho^2}{c^2 - b^2}, & \text{for } b < \rho < c, \\ 0, & \text{for } c < \rho. \end{cases} \quad (5.77)$$

Now, an elementary integration yields the magnetic energy per unit length of the cable:

$$\begin{aligned} \frac{U}{l} &= \frac{1}{2\mu_0} \int B^2 d^2r = \frac{\pi}{\mu_0} \int_0^\infty B^2 \rho d\rho = \frac{\mu_0 I^2}{4\pi} \left[\int_0^a \left(\frac{\rho}{a^2} \right)^2 \rho d\rho + \int_a^b \left(\frac{1}{\rho} \right)^2 \rho d\rho + \int_b^c \left(\frac{c^2 - \rho^2}{\rho(c^2 - b^2)} \right)^2 \rho d\rho \right] \\ &= \frac{\mu_0}{2\pi} \left[\ln \frac{b}{a} + \frac{c^2}{c^2 - b^2} \left(\frac{c^2}{c^2 - b^2} \ln \frac{c}{b} - \frac{1}{2} \right) \right] \frac{I^2}{2}. \end{aligned} \quad (5.78)$$

From here, and Eq. (72), we get the final answer:

$$\frac{L}{l} = \frac{\mu_0}{2\pi} \left[\ln \frac{b}{a} + \frac{c^2}{c^2 - b^2} \left(\frac{c^2}{c^2 - b^2} \ln \frac{c}{b} - \frac{1}{2} \right) \right]. \quad (5.79)$$

³² As a reminder, the mutual capacitance C between the conductors of such a system was calculated in Sec. 2.3. As will be discussed in Chapter 7 below, the pair of parameters L and C define the propagation of the most important, TEM mode of electromagnetic waves along the cable.

Note that for the particular case of a thin outer conductor, $c - b \ll b$, this expression reduces to

$$\frac{L}{l} \approx \frac{\mu_0}{2\pi} \left(\ln \frac{b}{a} + \frac{1}{4} \right), \quad (5.80)$$

where the first term in the parentheses may be traced back to the contribution of the magnetic field energy in the free space between the conductors. This distinction is important for some applications, because in superconductor cables, as well as resistive-metal cables at high frequencies (to be discussed in the next chapter), the field does not penetrate the conductor bulk, so that Eq. (80) is valid without the last term, $1/4$, in the parentheses, which is due to the magnetic field energy inside the wire.

As the last example, let us calculate the mutual inductance between a long straight wire and a round wire loop adjacent to it (Fig. 8), neglecting the thickness of both wires.

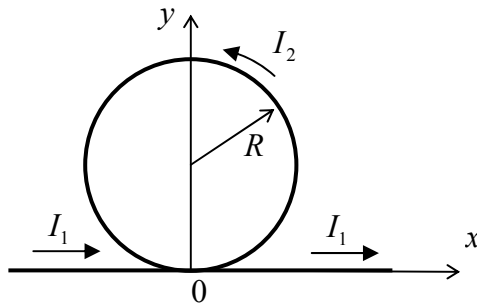


Fig. 5.8. Study case for the mutual inductance calculation.

Here there is no problem with using the last formalism, based on the magnetic flux calculation. Indeed, in the Cartesian coordinates shown in Fig. 8, Eq. (20) reads $B_1 = \mu_0 I_1 / 2\pi y$, giving the following magnetic flux through the round wire loop:

$$\Phi_{21} = \frac{\mu_0 I_1}{2\pi} \int_{-R}^{+R} dx \int_{R - (R^2 - x^2)^{1/2}}^{R + (R^2 - x^2)^{1/2}} dy \frac{1}{y} = \frac{\mu_0 I_1}{\pi} \int_0^R \ln \frac{R + (R^2 - x^2)^{1/2}}{R - (R^2 - x^2)^{1/2}} dx = \frac{\mu_0 I_1 R}{\pi} \int_0^1 \ln \frac{1 + (1 - \xi^2)^{1/2}}{1 - (1 - \xi^2)^{1/2}} d\xi. \quad (5.81)$$

This is a table integral equal to π ³³ so that $\Phi_{21} = \mu_0 I_1 R$, and the final answer for the mutual inductance $M = L_{12} = L_{21} = \Phi_{21}/I_1$ is finite (and very simple):

$$M = \mu_0 R, \quad (5.82)$$

despite magnetic field's divergence at the lowest point of the loop ($y = 0$). Note that in contrast with the finite *mutual* inductance of this system, *self*-inductances of both wires are formally infinite in the thin-wire limit – see, e.g., Eq. (80), that in the limit $b/a \gg 1$ describes a thin straight wire. However, since this divergence is very weak (logarithmic), it is quenched by any deviation from this perfect geometry. For example, a good estimate of the inductance of a wire of a large but finite length l may be obtained from Eq. (81) via the replacement of b with l :

$$L \sim \frac{\mu_0}{2\pi} l \ln \frac{l}{a}. \quad (5.83)$$

³³ See, e.g., MA Eq. (6.13) for $a = 1$.

(Note, however, that the exact result depends on where from/to the current flows beyond that segment.) A close estimate, with l replaced with $2\pi R$, and b replaced with R , is valid for the self-inductance of the round loop. A more exact calculation of this inductance, which would be asymptotically correct in the limit $a \ll R$, is a very useful exercise, which is highly recommended to the reader.³⁴

5.4. Magnetic dipole moment, and magnetic dipole media

The most natural way of description of magnetic media parallels that described in Chapter 3 for dielectrics, and is based on properties of *magnetic dipoles*. To introduce this notion quantitatively, let us consider, just as in Sec. 3.1, a spatially-localized system with current distribution $\mathbf{j}(\mathbf{r})$, whose magnetic field is measured at relatively large distances $r \gg r'$ (Fig. 9).

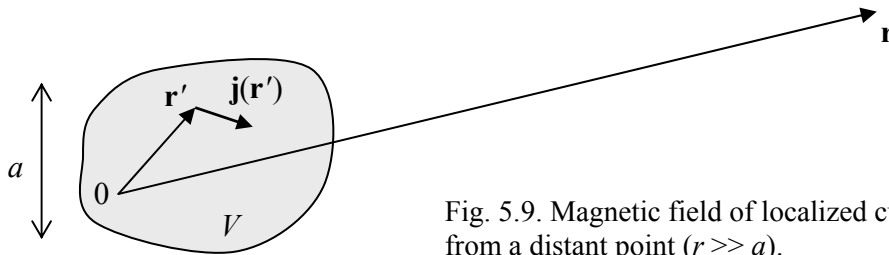


Fig. 5.9. Magnetic field of localized currents, observed from a distant point ($r \gg a$).

Applying the truncated Taylor expansion (3.4) to definition (28) of the vector potential, we get

$$\mathbf{A}(\mathbf{r}) \approx \frac{\mu_0}{4\pi} \left[\frac{1}{r} \int_V \mathbf{j}(\mathbf{r}') d^3 r' + \frac{1}{r^3} \int_V (\mathbf{r} \cdot \mathbf{r}') \mathbf{j}(\mathbf{r}') d^3 r' \right]. \quad (5.84)$$

Due to the vector character of this potential, we have to depart slightly from the approach of Sec. 3.1 and use the following vector algebra identity:³⁵

$$\int_V [f(\mathbf{j} \cdot \nabla g) + g(\mathbf{j} \cdot \nabla f)] d^3 r = 0 \quad (5.85)$$

that is valid for any pair of smooth (differentiable) scalar functions $f(\mathbf{r})$ and $g(\mathbf{r})$, and any vector function $\mathbf{j}(\mathbf{r})$ that, as the dc current density, satisfies the continuity condition $\nabla \cdot \mathbf{j} = 0$ and whose normal component vanishes on its surface.

First, let us use Eq. (85) with $f \equiv 1$ and g equal to any component of the radius-vector \mathbf{r} : $g = r_i$ ($i = 1, 2, 3$). Then it yields

$$\int_V (\mathbf{j} \cdot \mathbf{n}_i) d^3 r = \int_V j_i d^3 r = 0, \quad (5.86)$$

so that for the vector as the whole

³⁴ Its solution may be found, for example, just after Sec. 34 of L. Landau et al., *Electrodynamics of Continuous Media*, 2nd ed., Butterworth Heinemann, 1984.

³⁵ See, e.g., MA Eq. (12.3) with additional condition $j_n|_S = 0$, pertinent for space-restricted currents.

$$\int_V \mathbf{j}(\mathbf{r}) d^3r = 0, \quad (5.87)$$

showing that the first term in the right-hand part of Eq. (84) equals zero. Next, let us use Eq. (85) with $f = r_i$, $g = r_{i'}$ ($i, i' = 1, 2, 3$); then it yields

$$\int_V (r_i j_{i'} + r_{i'} j_i) d^3r = 0, \quad (5.88)$$

so that the i^{th} Cartesian component of the second integral in Eq. (84) may be transformed as

$$\begin{aligned} \int_V (\mathbf{r} \cdot \mathbf{r}') j_i d^3r' &= \int_V \sum_{i'=1}^3 r_i r'_{i'} j_i d^3r' = \frac{1}{2} \sum_{i'=1}^3 r_i \int_V (r'_{i'} j_i + r'_{i'} j_i) d^3r' \\ &= \frac{1}{2} \sum_{i'=1}^3 r_i \int_V (r'_{i'} j_i - r'_{i'} j_{i'}) d^3r' = -\frac{1}{2} \left[\mathbf{r} \times \int_V (\mathbf{r}' \times \mathbf{j}) d^3r' \right]_i. \end{aligned} \quad (5.89)$$

As a result, Eq. (85) may be rewritten as

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi} \frac{\mathbf{m} \times \mathbf{r}}{r^3}, \quad (5.90)$$

Magnetic
dipole and
its potential

where vector \mathbf{m} , defined as³⁶

$$\mathbf{m} \equiv \frac{1}{2} \int_V \mathbf{r} \times \mathbf{j}(\mathbf{r}) d^3r, \quad (5.91)$$

is called the *magnetic dipole moment* of our system - that itself, within approximation (90), is called the *magnetic dipole*.

Note a close analogy between \mathbf{m} and the angular momentum of a nonrelativistic particle with mass m_k :

$$\mathbf{L}_k \equiv \mathbf{r}_k \times \mathbf{p}_k = \mathbf{r}_k \times m_k \mathbf{v}_k, \quad (5.92)$$

where $\mathbf{p}_k = m_k \mathbf{v}_k$ is its mechanical momentum. Indeed, for a continuum of such particles with the same electric charge q , with the spatial density n , $\mathbf{j} = qn\mathbf{v}$, and Eq. (91) yields

$$\mathbf{m} = \int_V \frac{1}{2} \mathbf{r} \times \mathbf{j} d^3r = \int_V \frac{nq}{2} \mathbf{r} \times \mathbf{v} d^3r, \quad (5.93)$$

while the total angular momentum of such continuous system of particles of the same mass ($m_k = m_0$) is

$$\mathbf{L} = \int_V n m_0 \mathbf{r} \times \mathbf{v} d^3r,$$

so that we get a very straightforward relation

$$\mathbf{m} = \frac{q}{2m_0} \mathbf{L}. \quad (5.95)$$

\mathbf{m} vs. \mathbf{L}

³⁶ In the Gaussian units, definition (91) is kept valid, so that Eq. (90) is stripped of the factor $\mu_0/4\pi$.

For the orbital motion, this classical relation survives in quantum mechanics for operators and hence for eigenvalues, in whom the angular momentum is quantized in the units of the Plank's constant \hbar , so that for an electron, the orbital magnetic moment is always a multiple of the so-called *Bohr magneton*

$$\mu_B \equiv \frac{e\hbar}{2m_e}, \quad (5.96) \quad \text{Bohr magneton}$$

where m_e is the free electron mass.³⁷ However, for particles with spin, such a universal relation between vectors \mathbf{m} and \mathbf{L} is no longer valid. For example, electron's spin $s = 1/2$ gives contribution $\hbar/2$ to the mechanical momentum, but its contribution to the magnetic moment is still very close to μ_B .³⁸

The next important example of a magnetic dipole is a *planar* wire loop limiting area A (of an arbitrary shape), carrying current I , for which \mathbf{m} has a surprisingly simple form,

$$\mathbf{m} = I\mathbf{A}, \quad (5.97)$$

where the modulus of vector \mathbf{A} equals area A , and its direction is perpendicular to loop's plane. This formula may be readily proved by noticing that if we select the coordinate origin on the plane of the loop (Fig. 10), then the elementary component of the magnitude of integral (91),

$$m = \frac{1}{2} \left| \oint_C \mathbf{r} \times I d\mathbf{r} \right| = I \oint_C \left| \frac{1}{2} \mathbf{r} \times d\mathbf{r} \right| = I \oint_C \frac{1}{2} r^2 d\varphi, \quad (5.98)$$

is just the elementary area $dA = (1/2)r d\varphi = (1/2)r d(r \sin \varphi) = r^2 d\varphi/2$.

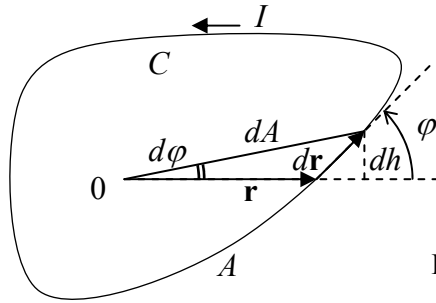


Fig. 5.10. Planar current loop.

The combination of Eqs. (96) and (97) allows a useful estimate of the scale of atomic currents, by finding what current I should flow in a circular loop of atomic size scale (the Bohr radius) $r_B \approx 0.5 \times 10^{-10}$ m, i.e. of area $A \approx 10^{-20}$ m², to produce a magnetic moment equal to μ_B .³⁹ The result is surprisingly macroscopic: $I \sim 1$ mA (quite comparable to the currents driving your earbuds :-). Though this estimate should not be taken too literally, due to the quantum-mechanical spread of electron's wavefunctions, it is very useful for getting a feeling how significant the atomic magnetism is and hence why ferromagnets may provide such a strong field.

³⁷ In SI units, $m_e \approx 0.91 \times 10^{-30}$ kg, so that $\mu_B \approx 0.93 \times 10^{-23}$ J/T.

³⁸ See, e.g., QM Sec. 4.1 and beyond.

³⁹ Another way to arrive at the same estimate is to take $I \sim ef = e\omega/2\pi$ with $\omega \sim 10^{16}$ s⁻¹ being the typical frequency of radiation due to atomic interlevel quantum transitions.

After these illustrations, let us return to Eq. (90). Plugging it into the general formula (27), we may calculate the magnetic field of a magnetic dipole:

Magnetic
dipole's
field

$$\mathbf{B}(\mathbf{r}) = \frac{\mu_0}{4\pi} \frac{3\mathbf{r}(\mathbf{r} \cdot \mathbf{m}) - \mathbf{m}r^2}{r^5}. \quad (5.99)$$

The structure of this formula *exactly* duplicates that of Eq. (3.15) for the electric dipole field. Because of this similarity, the energy of a dipole in an external field, and hence the torque and force exerted on it by the field, are also absolutely similar to the expressions for an electric dipole - see Eqs. (3.15)-(3.18):

Magnetic
dipole
in external
field

$$U = -\mathbf{m} \cdot \mathbf{B}_{\text{ext}}, \quad (5.100)$$

and as a result,

$$\boldsymbol{\tau} = \mathbf{m} \times \mathbf{B}_{\text{ext}}, \quad (5.101)$$

$$\mathbf{F} = \nabla(\mathbf{m} \cdot \mathbf{B}_{\text{ext}}). \quad (5.102)$$

Now let us consider a system of many magnetic dipoles (e.g., atoms or molecules), distributed in space with density n . Then we can use Eq. (90) (generalized in the evident way for an arbitrary position, \mathbf{r}' , of a dipole), and the linear superposition principle, to calculate the “macroscopic” component of the vector-potential \mathbf{A} - in other words, dipole's potential averaged over short-scale variations on the inter-dipole distances:

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{\mathbf{M}(\mathbf{r}') \times (\mathbf{r} - \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|^3} d^3r', \quad (5.103)$$

where $\mathbf{M} \equiv n\mathbf{m}$ is the *macroscopic* (average) *magnetization*, i.e. the magnetic moment per unit volume. Transforming this integral absolutely similarly to how Eq. (3.27) had been transformed into Eq. (3.29), we get:

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{\nabla' \times \mathbf{M}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3r'. \quad (5.104)$$

Comparing this result with Eq. (28), we see that $\nabla \times \mathbf{M}$ is equivalent, in its effect, to the density \mathbf{j}_{ef} of a certain effective “magnetization current”. Just as the electric-polarization “charge” ρ_{ef} discussed in Sec. 3.2 (see Fig. 3.3), $\mathbf{j}_{\text{ef}} = \nabla \times \mathbf{M}$ may be interpreted the uncompensated part of vortex currents representing single magnetic dipoles (Fig. 11).

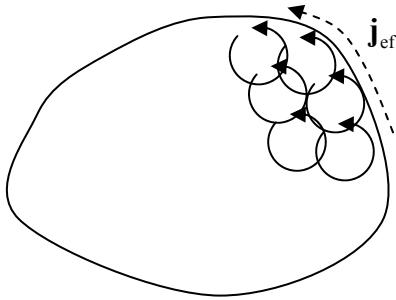


Fig. 5.11. Cartoon illustrating the physical nature of the “magnetization current” $\mathbf{j}_{\text{ef}} = \nabla \times \mathbf{M}$.

Now, using Eq. (28) to add the possible contribution from “stand-alone” currents \mathbf{j} , not included into the currents of microscopic dipoles, we get the general equation for the vector-potential of the macroscopic field:

$$\mathbf{A}(\mathbf{r}) = \frac{\mu_0}{4\pi} \int \frac{[\mathbf{j}(\mathbf{r}') + \nabla' \times \mathbf{M}(\mathbf{r}')] }{|\mathbf{r} - \mathbf{r}'|} d^3 r'. \quad (5.105)$$

Repeating the calculations that have led us from Eq. (28) to the Maxwell equation (35), with the account of the magnetization current term, for the macroscopic magnetic field \mathbf{B} we get⁴⁰

$$\nabla \times \mathbf{B} = \mu_0 (\mathbf{j} + \nabla \times \mathbf{M}). \quad (5.106)$$

Following the same philosophy as in Sec. 3.2, we may recast this equation as

$$\nabla \times \mathbf{H} = \mathbf{j}, \quad (5.107)$$

where a new field defined as

$$\mathbf{H} \equiv \frac{\mathbf{B}}{\mu_0} - \mathbf{M},$$

(5.108) Magnetic field \mathbf{H}

by historic reasons (and very unfortunately) is also called the *magnetic field*.⁴¹ It is crucial to remember that the physical sense of field \mathbf{H} is very much different from field \mathbf{B} . In order to understand the difference better, let us use Eq. (107) to complete a macroscopic analog of system (36), called the *macroscopic Maxwell equations* (again, so far for the stationary case $\partial/\partial t = 0$):

$$\begin{aligned} \nabla \times \mathbf{E} &= 0, & \nabla \times \mathbf{H} &= \mathbf{j}, \\ \nabla \cdot \mathbf{D} &= \rho, & \nabla \cdot \mathbf{B} &= 0. \end{aligned} \quad (5.109)$$

Stationary macroscopic Maxwell equations

One can clearly see that the roles of vector fields \mathbf{D} and \mathbf{H} are very similar: they could be called “would-be” fields - which *would be* induced by stand-alone charges and currents, if the media had not modified them by its dielectric and/or magnetic polarization.

⁴⁰ Similarly to the situation with the electric dipoles (see Eq. (3.24) and its discussion), it may be shown that the magnetic field of any closed current loop (or any system of such loops) satisfies the following equality:

$$\int_{r < R} \mathbf{B}(\mathbf{r}) d^3 r = (2/3) \mu_0 \mathbf{m},$$

where the integral is over any sphere confining all the currents. On the other hand, for field (99), derived from the asymptotic approximation (90), such integral vanishes. In order to get a course-grain description of the magnetic field of a small system located at $r = 0$, which would be valid everywhere (though at $r \sim a$, only approximately), Eq. (99) should be modified as follows:

$$\mathbf{B}_{cg}(\mathbf{r}) = \frac{\mu_0}{4\pi} \left(\frac{3\mathbf{r}(\mathbf{r} \cdot \mathbf{m}) - \mathbf{m}r^2}{r^5} + \frac{8\pi}{3} \mathbf{m} \delta(\mathbf{r}) \right).$$

Hence, strictly speaking, the macroscopic field \mathbf{B} participating in Eq. (106) and beyond is the average *long-range* field of the magnetic dipoles (plus of the stand-alone currents \mathbf{j}) rather than the genuine average magnetic field.

⁴¹ This confusion is exacerbated by the fact that in Gaussian units, Eq. (108) has the form $\mathbf{H} = \mathbf{B} - 4\pi\mathbf{M}$, and hence fields \mathbf{B} and \mathbf{H} has one dimensionality (and are equal in free space!) - though the unit of \mathbf{H} has a different name (*oersted*, abbreviated as Oe). Mercifully, in the SI units, the dimensionality of \mathbf{B} and \mathbf{H} is different, with the unit of \mathbf{H} being called *ampere per meter*.

Despite this similarity, let me note an important difference of signs in the relation (3.33) between \mathbf{E} , \mathbf{D} , and \mathbf{P} , on one hand, and relation (108) between \mathbf{B} , \mathbf{H} , and \mathbf{M} , on the other hand. It is *not* just the matter of definition. Indeed, due to the similarity of Eqs. (3.15), and (100), including similar signs, the electric and magnetic fields both try to orient the corresponding dipole moments along the field. Hence, in the media that allow such orientation (and as we will see momentarily, for magnetic media it is not always the case), the induced polarizations \mathbf{P} and \mathbf{M} are directed along, respectively, vectors \mathbf{E} and \mathbf{B} . According to Eq. (3.33), if the would-be field \mathbf{D} is fixed - say, by a fixed stand-alone charge distribution $\rho(\mathbf{r})$ - such polarization *reduces* the genuine average electric field $\mathbf{E} = (\mathbf{D} - \mathbf{P})/\epsilon_0$. On the other hand, Eq. (108) shows that in a magnetic media with fixed would-be field \mathbf{H} , magnetic polarization with $\mathbf{M} \uparrow \mathbf{B}$ *enhances* the average magnetic field $\mathbf{B} = (\mathbf{H} + \mathbf{M})/\mu_0$. This difference may be traced back to the sign difference in the initial relations (1.1) and (5.1), i.e. to the basic fact that charges of the same sign repulse, while currents of the same direction attract each other.

In order to form a complete system of differential equations, the macroscopic Maxwell equations (109) have to be complemented with “material relations” $\mathbf{D} \leftrightarrow \mathbf{E}$, $\mathbf{j} \leftrightarrow \mathbf{E}$, and $\mathbf{B} \leftrightarrow \mathbf{H}$. In previous two chapters we already discussed, in brief, two of them; let us proceed to the last one.

5.5. Magnetic materials

A major difference between the dielectric and magnetic material equations $\mathbf{D}(\mathbf{E})$ and $\mathbf{B}(\mathbf{H})$ is that while a typical dielectric media *reduces* the external electric field, magnetic media may *either reduce or enhance* it. In order to quantify this fact, let us consider the so-called *linear magnetics* in which \mathbf{M} (and hence \mathbf{B}) are proportional to \mathbf{H} . Just as in dielectrics, in material without spontaneous magnetization, such linearity at relatively low fields follows from the Taylor expansion of function $\mathbf{M}(\mathbf{B})$. For isotropic materials, this proportionality is characterized by a scalar - either the *magnetic permeability* μ , defined by the following relation:

Magnetic permeability

$$\mathbf{B} \equiv \mu \mathbf{H}, \quad (5.110)$$

or the *magnetic susceptibility*⁴² defined as

Magnetic susceptibility

$$\mathbf{M} = \chi_m \mathbf{H}. \quad (5.111)$$

Plugging these relations into Eq. (108), we see that these two parameters are not independent, but are related as

χ_m vs. μ

$$\mu = (1 + \chi_m) \mu_0. \quad (5.112)$$

Note that despite the superficial similarity between Eqs. (110)-(111) and relations (3.35)-(3.38) for linear dielectrics:

⁴² According to Eq. (110) (i.e. in SI units), χ_m is dimensionless, while μ has the same the same dimensionality as μ_0 . In the Gaussian units, μ is dimensionless, $(\mu)_{\text{Gaussian}} = (\mu)_{\text{SI}}/\mu_0$, and χ_m is also introduced differently, as $\mu = 1 + 4\pi\chi_m$. Hence, just as for the electric susceptibilities, these dimensionless coefficients are different in the two systems: $(\chi_m)_{\text{SI}} = 4\pi(\chi_m)_{\text{Gaussian}}$. Note also that χ_m is formally called the *volume* magnetic susceptibility, in order to distinguish it from the *molecular* susceptibility χ defined by a similar relation, $\mathbf{m} \equiv \chi \mathbf{H}$, where \mathbf{m} is the average induced magnetic moment of a single dipole – e.g., a molecule. Evidently, in a dilute medium, i.e. in the absence of substantial dipole-dipole interaction, $\chi_m = n\chi$, where n is the dipole density.

$$\mathbf{D} = \varepsilon \mathbf{E}, \quad \mathbf{P} = \chi_e \varepsilon_0 \mathbf{E}, \quad \varepsilon = (1 + \chi_e) \varepsilon_0, \quad (5.113)$$

there is an important conceptual difference between them. Namely, while vector \mathbf{E} in the right-hand parts of Eqs. (113) is the real (average) electric field, vector \mathbf{H} in the right-hand part of Eqs. (110)-(111) represents a “would-be” magnetic field, in all aspects similar to vector \mathbf{D} rather than \mathbf{E} . For relatively dense media, whose polarization may affect the genuine fields substantially, this difference between parameters ε and μ may make their properties (e.g., the Kramers-Kronig relations, to be discussed in Sec. 7.3) rather different.

Another difference between parameters ε and μ (and hence between χ_e and χ_m) is evident from Table 1 which lists the values of magnetic susceptibility for several materials. It shows that in contrast to linear dielectrics whose susceptibility χ_e is always positive, i.e. the dielectric constant $\varepsilon_r = \chi_e + 1$ is always larger than 1 (see Table 3.1), linear magnetics may be either *paramagnets* ($\chi_m > 0$, i. e. $\mu > \mu_0$) or *diamagnets* ($\chi_m < 0$, $\mu < \mu_0$).

Table 5.1. Magnetic susceptibility (χ_m)_{SI} of a few representative (and/or important) materials^(a)

“Mu-metal” (75% Ni + 15% Fe + a few %% of Cu and Mo)	~20,000 ^(b)
Permalloy (80% Ni + 20% Fe)	~8,000 ^(b)
“Soft” (or “transformer”) iron	~4,000 ^(b)
Nickel	~100
Aluminum	+2×10 ⁻⁵
Diamond	-2×10 ⁻⁵
Copper	-7×10 ⁻⁵
Water	-9×10 ⁻⁶
Bismuth (the strongest non-superconducting diamagnet)	-1.7×10 ⁻⁴

^(a)The table does not include bulk superconductors, which in a crude (“macroscopic”) approximation may be described as perfect diamagnets (with $\mathbf{B} = 0$, i.e. $\chi_m = -1$ and $\mu = 0$), though the actual physics of this phenomenon is more complex – see Sec. 6.3 below.

^(b) The exact values of χ_m for soft ferromagnetic materials depend not only on their exact composition, but also on their thermal processing (“annealing”). Moreover, due to unintentional vibrations, the extremely high χ_m of such materials may somewhat decay with time, though may be restored to approach the original value by new annealing.

The reason of this difference is that in dielectrics, two different polarization mechanisms (schematically illustrated by Fig. 12) lead to the same sign of the average polarization. The first of them takes place in atoms without their own spontaneous polarization. A crude classical image of such an atom is an isotropic cloud of negatively charged electrons surrounding a positively charged nucleus - see Fig. 12a. The external electric field shifts the positive charge in the direction of \mathbf{E} , and negative charges in the opposite direction, thus creating a dipole with aligned vectors \mathbf{p} and \mathbf{E} , and hence positive

polarizability α_{mol} - see Eq. (3.39). As a result, the electric susceptibility is also positive – see Eqs. (3.41) or (3.71).

In the second case (Fig. 12b) of a gas or liquid consisting of *polar molecules*, each molecule has its own, spontaneous dipole moment \mathbf{p}_0 even in the absence of external electric field. (A typical example is a water molecule H_2O , with the positive oxygen ion positioned out of the line connecting two positive hydrogen atoms, thus producing a spontaneous dipole with moment's magnitude $p_0 \approx e \times 0.38 \times 10^{-10} \text{ m}$.) However, in the absence of the applied electric field, the orientation of such dipoles is random, so that the average polarization $\mathbf{P} = n\langle\mathbf{p}_0\rangle$ equals zero. A weak applied field does not change the magnitude of the dipole moments significantly, but creates their preferential orientation along the field (in order to decrease the potential energy $U = -\mathbf{p}_0 \cdot \mathbf{E}$), thus creating a nonvanishing vector average $\langle\mathbf{p}_0\rangle$ directed along \mathbf{E} . If the applied field is not too high ($p_0 E \ll k_B T$), the induced polarization $\mathbf{P} = n\langle\mathbf{p}_0\rangle$ is proportional to \mathbf{E} , again giving a positive polarizability α_{mol} .⁴³

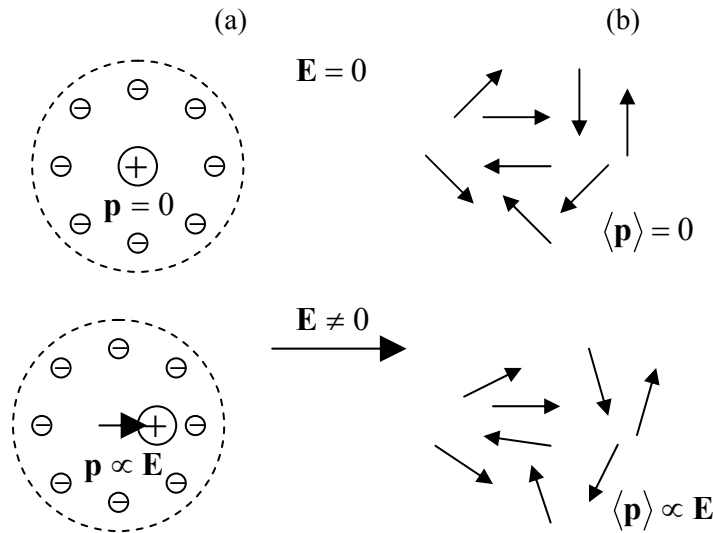


Fig. 5.12. Cartoons of two types of induced electrical polarization: (a) elementary dipole induction and (b) partial ordering of spontaneous elementary dipoles.

Returning to magnetics, the second of the above mechanisms, i.e. the ordering of spontaneous dipoles by the applied field, is responsible for the paramagnetism. Again, now according to Eq. (100), such field tends to align the dipoles along its direction, so that the average direction of spontaneous elementary moments \mathbf{m}_0 , and hence the direction of \mathbf{M} , is the same as that of the average field \mathbf{B} (i.e., for a diluted media, of $\mathbf{H} \approx \mathbf{B}/\mu_0$), resulting in a positive susceptibility χ_m . However, in contrast to the electric polarization, there is a mechanism of magnetic polarization, called the *orbital* (or “Larmor”⁴⁴) *diamagnetism*, which gives $\chi_m < 0$. As its simplest model, let us consider the orbital motion of an atomic electron as classical particle of mass m_0 , with electric charge q , about an immobile attractive center - modeling the atomic nucleus. As classical mechanics tells us, the central attractive force does

⁴³ The proportionality of $|\langle p_0 \rangle|$ (and hence P) to E is a result of a dynamic balance between the dipole-orienting torque (101) and disordering thermal fluctuations. A qualitative description of such balances is one of the main tasks of statistical mechanics - see, e.g., SM Chapters 2 and 4. However, the very fact of proportionality $P \propto E$ in low fields may be readily understood as the result of the Taylor expansion of function $P(E)$ at $E \rightarrow 0$.

⁴⁴ After J. Larmor (1857 – 1947) who first described the torque-induced precession mathematically.

not change particle's angular momentum $\mathbf{L} \equiv m_0 \mathbf{r} \times \mathbf{v}$, but the applied magnetic field \mathbf{B} (that may be taken uniform on the atomic scale) does, due to the torque (101) it applies to magnetic moment (95):

$$\frac{d\mathbf{L}}{dt} = \boldsymbol{\tau} = \mathbf{m} \times \mathbf{B} = \frac{q}{2m_0} \mathbf{L} \times \mathbf{B}. \quad (5.114)$$

The vector diagram in Fig. 13 shows that in the limit of relatively weak field, when the magnitude of the angular momentum \mathbf{L} may be considered constant, this equation describes the rotation (called the *torque-induced precession*⁴⁵) of vector \mathbf{L} about the direction of vector \mathbf{B} , with angular frequency $\boldsymbol{\Omega} = -q\mathbf{B}/2m_0$, independent on angle θ . Let me leave for the reader to use Eq. (114) for checking that, irrespectively the sign of charge q , the resulting additional magnetic moment $\Delta\mathbf{m}$ has a direction *opposite* to that of vector \mathbf{B} , and hence χ_m is negative, leading to the Larmor diamagnetism.⁴⁶

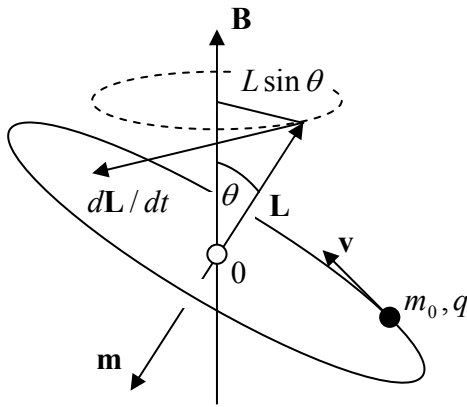


Fig. 5.13. Torque-induced precession of a charged particle in a magnetic field.

An important conceptual question is what exactly prevents the initial magnetic moment \mathbf{m} that, according to Eq. (95), is associated with the angular momentum \mathbf{L} of the electron, from turning along the magnetic field, just as in the second polarization mechanism illustrated by Fig. 12b - thus decreasing the potential energy (100) of the system. The answer is the same as for the usual mechanical top – it “wants” to fall due to the gravity field, but cannot do that due to the mechanical inertia. In classical physics, even a small friction (dissipation) eventually drains top’s rotational kinetic energy, and it falls. However, in quantum mechanics the ground-state “motion” of electrons in an atom is not subjected to friction, because they cannot be brought to full rest due to Heisenberg’s uncertainty principle. Somewhat counter-intuitively, the magnetic moments due to such fully-quantum effect as spin are much more susceptible to interaction with environment, so that in atoms with uncompensated spins, the magnetic dipole orientation mechanism prevails over the orbital diamagnetism, and the materials incorporating such atoms usually exhibit net paramagnetism – see Table 1.

Due to possible strong interactions between elementary dipoles, magnetism of materials is an extremely rich field of physics, with numerous interesting phenomena and elaborated theories.

⁴⁵ For a detailed discussion of the effect see, e.g., CM Sec. 6.5.

⁴⁶ The quantum-mechanical treatment (see, e.g., QM Sec. 6.4) confirms this qualitative picture, while giving quantitative corrections to the classical result for χ_m .

Unfortunately, all this physics is well outside the framework of this course, and I have to refer the interested reader to special literature,⁴⁷ but still need to mention its key notions.

Most importantly, a sufficiently strong dipole-dipole interaction may lead to their spontaneous ordering, even in the absence of the applied field. This ordering may correspond to either parallel alignment of the atomic dipoles (*ferromagnetism*) or anti-parallel alignment of the adjacent dipoles (*antiferromagnetism*). Evidently, the external effects of ferromagnetism are stronger, because such phase corresponds to a substantial spontaneous magnetization \mathbf{M} . (This value is frequently called the *saturation magnetization*, \mathbf{M}_s , while the corresponding magnitude of $\mathbf{B} = \mu_0 \mathbf{M}$ is called either the *saturation magnetic field*, or the *remanence field*, \mathbf{B}_R). The direction of \mathbf{B}_R may be switched by the application of an external magnetic field, with a magnitude above certain value H_C called *coercivity*,⁴⁸ leading to the well-known hysteretic loops on the $[B, H]$ plane - see Fig. 14 for a typical example.

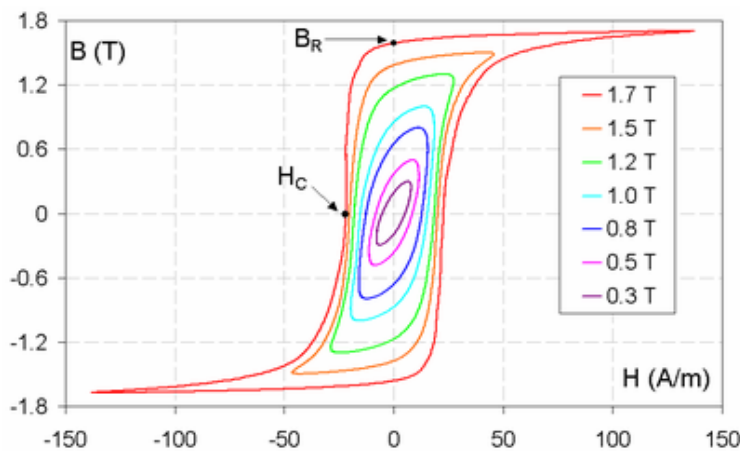


Fig. 5.14. Experimental magnetization curves of specially processed (cold-rolled) transformer steel, i.e. a solid solution of $\sim 10\%$ C and $\sim 6\%$ Si in Fe. (Adapted from www.thefullwiki.org/Hysteresis.)

In relatively low fields, $H \ll H_C$, such materials may be described as *hard* (or “permanent”) *ferromagnets*; at such approximate treatment, magnetization \mathbf{M} is considered constant. On the other hand, the theory needed for a fair description of phenomena at $H \sim H_C$ is rather complicated. Indeed, the direction of magnetization of crystals may be affected by the anisotropy of the crystal lattice. Because of that, typical non-crystalline ferromagnetic materials (like steel, permalloy, “mu-metal”, etc.) consist of randomly oriented *magnetic domains*, each with certain spontaneous magnetization direction. The magnetic interaction of the domain with its neighbors and the external field determines the evolution of its magnetization and hence the average magnetic properties of the ferromagnet. In particular, such interaction explains why the hysteresis loop shape is dependent on the cycled field amplitude and cycling history – see Fig. 14. A very important class of multi-domain materials is the so-called *soft ferromagnets*, whose coercivity is relatively low. At low cycled field amplitude, the soft ferromagnets behave, on the average, as linear magnetics with very high values of χ_m and hence μ (see the top rows of Table 1, and Fig. 14) that are highly dependent on the material’s fabrication technology and its post-fabrication thermal and mechanical treatments.

⁴⁷ See, e.g., D. J. Jiles, *Introduction to Magnetism and Magnetic Materials*, 2nd ed., CRC Press, 1998, or R. C. O’Handley, *Modern Magnetic Materials*, Wiley, 1999.

⁴⁸ Materials with very high coercivity H_C are frequently called *hard ferromagnets* or *permanent magnets*.

High values of χ_m are also pertinent to magnetics in which the molecular dipole interaction is relatively weak, so that their ferromagnetic ordering may be destroyed by thermal fluctuations, if temperature is increased above the so-called *Curie temperature* T_C . At $T > T_C$, such materials behave as paramagnets, with susceptibility obeying the *Curie-Weiss law*

$$\chi_m \propto \frac{1}{T - T_C}. \quad (5.115)$$

(At vanishing moment interaction, $T_C \rightarrow 0$, and Eq. (115) is reduced to the *Curie law* $\chi_m \propto 1/T$ typical for weak paramagnets.) The transition between the ferromagnetic and paramagnetic phase at $T = T_C$ is the classical example of *continuous phase transitions*, similar to that between the paraelectric and ferroelectric phases of a dielectric. In both cases, the “macroscopic” (average) polarization – either \mathbf{M} or \mathbf{P} – plays the role of the so-called *order parameter* that (in the absence of external fields) appears at $T = T_C$ and increases gradually at the further reduction of temperature.⁴⁹

Before returning to magnetostatics per se, I have to mention the large practical role played by hard ferromagnetic materials (well beyond refrigerator magnets :-). Indeed, despite the decades of the exponential (*Moore's-law*) progress of semiconductor electronics, most computer data storage systems are still based on the *hard disk drives* whose active medium is a submicron-thin ferromagnetic layer, with bits stored in the form of the direction of the spontaneous magnetization of small film spots. This technology has reached a fantastic sophistication,⁵⁰ with recording data density approaching 10^{12} bits per square inch. Only recently it has started to be seriously challenged by the so-called *solid state drives* based on the flash semiconductor memories already mentioned in Chapter 3.

5.6. Systems with magnetics

Similarly to the electrostatics of linear dielectrics, magnetostatics of *linear* magnetics is very simple in the particular case when the stand-alone currents are deeply embedded into a medium with a constant permeability μ . Indeed, in this case, boundary conditions on the distant surface of the media do not affect the solution of the boundary problem described by the magnetic equations of the macroscopic Maxwell system (109). Now let us assume that we know the solution $\mathbf{B}_0(\mathbf{r})$ of the magnetic pair of the genuine (“microscopic”) Maxwell equations (36) in free space, i.e. when the genuine current density \mathbf{j} coincides with that of stand-alone currents. Then the macroscopic equations and the material equation (110) are completely satisfied with the pair of functions

$$\mathbf{H}(\mathbf{r}) = \frac{\mathbf{B}_0(\mathbf{r})}{\mu_0}, \quad \mathbf{B}(\mathbf{r}) = \mu \mathbf{H}(\mathbf{r}) = \frac{\mu}{\mu_0} \mathbf{B}_0(\mathbf{r}). \quad (5.116)$$

Hence the only effect of a complete filling a system of fixed currents with a uniform, linear magnetic is the *increase* of the magnetic field \mathbf{B} at all points by the same constant factor $\mu/\mu_0 \equiv 1 + \chi_m$. (As a reminder, a similar filling of a system of fixed charges with a uniform, linear dielectric leads to a *reduction* of the electric field \mathbf{E} by factor $\varepsilon/\varepsilon_0 = \varepsilon_r = 1 + \chi_e$.)

⁴⁹ A discussion of such transitions may be found, in particular, in SM Chapter 4.

⁵⁰ “A magnetic head slider [the read/write head – KKL] flying over a [rather uneven – KKL] disk surface with a flying height of 25 nm with a relative speed of 20 meters/second is equivalent to an aircraft flying at a physical spacing of 0.2 μm at 900 kilometers/hour.” B. Bhushan, as quoted in a (generally good) book by G. Hadjipanayis, *Magnetic Storage Systems Beyond 2000*, Springer, 2001.

However, this simple result is generally invalid in the case of non-uniform (or piece-wise uniform) magnetic samples. Theoretical analyses of magnetic field distribution in such non-uniform systems may be facilitated by two additional tools. First, integrating the macroscopic Maxwell equation (107) along a closed contour C limiting a smooth surface S , and using the Stokes theorem, we get the macroscopic version of the Ampère law (37):

Macroscopic
Ampère
law

$$\oint_C \mathbf{H} \cdot d\mathbf{r} = I. \quad (5.117)$$

This is exactly the replica of the “microscopic” equation Eq. (37), with the replacement $\mathbf{B}/\mu_0 \rightarrow \mathbf{H}$.

Let us apply this relation to a boundary between two regions with constant, but different μ , with no stand-alone currents on the border, similarly how this was done for field \mathbf{E} in Sec. 3.4 - see Fig. 3.5. The result is similar as well:

$$H_\tau = \text{const}. \quad (5.118)$$

On the other hand, the integration of the Maxwell equation (29) over a Gaussian pillbox enclosing a border fragment (again similar to that shown in Fig. 3.5) yields the result similar to Eq. (3.46):

$$B_n = \text{const}, \quad \text{i.e. } \mu H_n = \text{const}. \quad (5.119)$$

Let us use these boundary conditions, first, to see what happens with a thin sheet of magnetic material (or any other strongly elongated sample) placed *parallel* to a uniform external field \mathbf{H}_0 . Such sample cannot noticeably disturb the field in the free space outside it: $\mathbf{H}_{\text{ext}} = \mathbf{H}_0$, $\mathbf{B}_{\text{ext}} = \mathbf{H}_{\text{ext}}/\mu_0 = \mathbf{H}_0/\mu_0$. Now applying Eq. (118) to the dominating, large-area interfaces, we get $\mathbf{H}_{\text{int}} = \mathbf{H}_0$, i.e., $\mathbf{B}_{\text{int}} = (\mu/\mu_0) \mathbf{B}_0$.⁵¹ The fact of constancy of field \mathbf{H} in this geometry explains why this field is used as the horizontal axis in plots like Fig. 14: such measurements are typically carried out by placing an elongated sample of the material into the uniform field – say the one produced by a long solenoid.

Samples of other geometries may create strong perturbations of the external field, extended to distances of the order of the transversal dimensions of the sample. In order to analyze such problems, we may benefit from a simple, partial differential equation for a scalar function, e.g., the Laplace equation, because in Chapter 2 we have learned how to solve it for many simple geometries. In magnetostatics, the introduction of a scalar potential is generally impossible due to the vortex-like magnetic field lines, but if there are no stand-alone currents within the region we are interested in, then the Maxwell equation (32) for field \mathbf{H} is reduced to $\nabla \times \mathbf{H} = 0$, and we may introduce the scalar potential of the magnetic field, ϕ_m , using the relation similar to Eq. (1.33):

$$\mathbf{H} = -\nabla \phi_m. \quad (5.120)$$

Combining it with the homogenous Maxwell equation for magnetic field, $\nabla \cdot \mathbf{B} = 0$, we arrive at the familiar differential equation,

$$\nabla \cdot (\mu \nabla \phi_m) = 0, \quad (5.121)$$

that, for a uniform media ($\mu = \text{const}$), is reduced to our beloved Laplace equation. Moreover, Eqs. (118) and (119) give the very familiar boundary conditions: first

⁵¹ The reader is highly encouraged to carry out a similar analysis of fields inside narrow gaps cut in a linear magnetic, similar to that carried out for linear dielectrics in Sec. 3.3 – see Fig. 3.6 and its discussion.

$$\frac{\partial \phi_m}{\partial \tau} = \text{const}, \quad (5.122a)$$

which is equivalent to

$$\phi_m = \text{const}, \quad (5.122b)$$

and also

$$\mu \frac{\partial \phi_m}{\partial n} = \text{const}. \quad (5.123)$$

Note that these boundary conditions are similar for (3.46) and (3.47) of electrostatics, with the replacement $\varepsilon \rightarrow \mu$.⁵²

Let us analyze the geometric effects on magnetization, using the (too?) familiar structure: a sphere, made of a linear magnetic material, in a uniform external field. Since the differential equation and boundary conditions are similar to those of the similar electrostatics problem (see Fig. 3.8), we can use the above analogy to recycle the solution we already have got – see Eqs. (3.55)-(3.56). Just as in the electric case, the field outside the sphere, with potential

$$(\phi_m)_{r>R} = H_0 \left(-r + \frac{\mu - \mu_0}{\mu + 2\mu_0} \frac{R^3}{r^2} \right) \cos \theta, \quad (5.125a)$$

is a sum of the uniform external field \mathbf{H}_0 and the dipole field (99) with the following induced magnetic dipole moment of the sphere:⁵³

$$\mathbf{m} = 4\pi \frac{\mu - \mu_0}{\mu + 2\mu_0} R^3 \mathbf{H}_0. \quad (5.125b)$$

On the contrary, the internal field is perfectly uniform:

$$(\phi_m)_{r<R} = -H_0 \frac{3\mu_0}{\mu + 2\mu_0} r \cos \theta, \quad \frac{H_{\text{int}}}{H_0} = \frac{3\mu_0}{\mu + 2\mu_0}, \quad \frac{B_{\text{int}}}{B_0} = \frac{\mu H_{\text{int}}}{\mu_0 H_0} = \frac{3\mu}{\mu + 2\mu_0}. \quad (5.126)$$

Note that though \mathbf{H} inside the sphere is not equal to its value of the external field \mathbf{H}_0 . This example shows that the interpretation of \mathbf{H} as the “would-be” magnetic field generated by external

⁵² This similarity may seem strange, because earlier we have seen that parameter μ is physically more similar to $1/\varepsilon$. The reason for this paradox is that in magnetostatics, the introduced potential ϕ_m is traditionally used to describe the “would-be field” \mathbf{H} , while in electrostatics, potential ϕ describes the real (average) electric field \mathbf{E} . (This tradition persists from the old days when \mathbf{H} was perceived as a genuine magnetic field.)

⁵³ Instead of differentiating the ϕ_m given by Eq. (125a), we may use the absolute similarity of Eqs. (3.13) and (99), to derive from Eq. (3.17) a similar expression for the magnetic potential of an arbitrary magnetic dipole:

$$\phi_m = \frac{1}{4\pi} \frac{m \cos \theta}{r^2}.$$

Now comparing this formula with the second term of Eq. (125a), we immediately get Eq. (125b).

currents \mathbf{j} should not be exaggerated into saying that its distribution is independent on the magnetic bodies in the system.⁵⁴

In the limit $\mu \gg \mu_0$, Eqs. (126) yield $H_{\text{int}}/H_0 \ll 1$, $B_{\text{int}}/H_0 = 3\mu_0$, the factor 3 being specific for the particular geometry of the sphere. If a sample is stretched along the applied field, this limitation of the field concentration is gradually removed, and B_{int} tends to its maximum value $\mu H_0 \gg B_{\text{ext}}$, as was discussed above. This effect of “magnetic line concentration” in high- μ materials is used in such practically important devices as transformers, in which two multi-turn coils are wound on a ring-shaped (e.g., toroidal, see Fig. 6b) core made of a soft ferromagnetic material (such as the *transformer steel*, see Table 1) with $\mu \gg \mu_0$. This minimizes the number of “stray” field lines, and makes the magnetic flux Φ piercing each wire turn (of either coil) virtually the same – the equality important for secondary voltage induction – see the next chapter.

The second theoretical tool, frequently useful for problem solution, is a macroscopic expression for magnetic field energy U . For a system with linear magnetic materials, we may repeat the transformation of Eq. (55), made in Sec. 3, but with due respect to the magnetization, i.e. replacing \mathbf{j} not from Eq. (56), but from Eq. (107). As a result, instead of Eq. (57) we get

$$U = \int_V u(\mathbf{r}) d^3r, \quad \text{with } u = \frac{\mathbf{B} \cdot \mathbf{H}}{2} = \frac{B^2}{2\mu} = \frac{\mu H^2}{2}, \quad (5.127)$$

This result is evidently similar to Eq. (3.79) of electrostatics.

For the general case of nonlinear magnetics, calculations similar to those resulting in Eq. (3.82) give the following analog of that relation:

$$\delta u = \mathbf{H} \cdot \delta \mathbf{B}, \quad (5.128)$$

for a linear magnetic yielding Eq. (127). Similarly to the electrostatics of dielectrics, we may argue that according to Eq. (128), in systems with magnetic media, \mathbf{H} plays the role of the generalized force, and \mathbf{B} of the generalized coordinate (per unit volume).⁵⁵ As the result, the Gibbs potential energy, whose minimum corresponds to the stable equilibrium of the system in an external field \mathbf{H}_{ext} , is

$$\mathcal{G} = \int_V g(\mathbf{r}) d^3r, \quad \text{with } g(\mathbf{r}) \equiv u(\mathbf{r}) - \mathbf{H}_{\text{ext}} \cdot \mathbf{B}, \quad (5.129)$$

the expression to be compared with Eq. (3.84). Similarly, for a system with linear magnetics, the latter of these expressions may be integrated over the variations to give

⁵⁴ From the standpoint of mathematics, this happens because the solution to a boundary problem is determined by not only the differential equation inside the system (in our case, the Laplace equation for potential ϕ_m), but also by boundary conditions – which are affected by magnetics – see Eqs. (118)-(119).

⁵⁵ Note that in this respect, the analogy with electrostatics is incomplete. Indeed, according to Eq. (3.82), in electrostatics the role of a generalized coordinate is played by would-be field \mathbf{D} , and that of the generalized force, by the real (average) electric field \mathbf{E} . This difference may be traced back to the fact that electric field \mathbf{E} may perform work on a moving charged particle, while the magnetic part of the Lorentz force (10), $\mathbf{v} \times \mathbf{B}$, is always perpendicular to particle’s velocity, and its work equals zero. However, this difference does not affect the full analogy of expressions (3.79) and (127) for field energy density in linear media.

$$g(\mathbf{r}) = \frac{1}{2\mu} \mathbf{B} \cdot \mathbf{B} - \mathbf{H}_{\text{ext}} \cdot \mathbf{B} = \frac{1}{2\mu} (\mathbf{B} - \mu \mathbf{H}_{\text{ext}})^2 + \text{const}, \quad (5.130)$$

with similar consequences for the external magnetic field penetration into a system with magnetics. As a sanity check, for a uniform system with negligible fringe fields, such as a long solenoid filled with a uniform, linear magnetic material, Eq. (130) may be readily integrated over the sample volume to give

$$\mathcal{G}(\mathbf{r}) = \frac{1}{2\mu} (\mathbf{B} - \mu \mathbf{H}_{\text{ext}})^2 V + \text{const}, \quad (5.131)$$

so that the minimum of the Gibbs potential energy, i.e. the stable equilibrium of the system, corresponds to the result that has already been derived in the beginning of this section: $\mathbf{B} = \mu \mathbf{H}_{\text{ext}}$, i.e. $\mathbf{H} = \mathbf{H}_{\text{ext}}$.

For the important particular case of a long solenoid (Fig. 6a) filled with a linear magnetic material, we may find field H from Eq. (117), just as we used Eq. (37) in Sec. 2 for finding B for a similar empty solenoid, getting

$$H = In, \text{ and hence } B = \mu In. \quad (5.132)$$

Now we may plug this result into Eq. (127) to calculate the magnetic energy stored in the solenoid:

$$U = uV = \frac{\mu H^2}{2} lA = \frac{\mu (nI)^2 lA}{2}, \quad (5.132)$$

and then use Eq. (72) to calculate its self-inductance:

$$L = \frac{U}{I^2/2} = \mu n^2 lA \quad (5.133)$$

- as evident generalization of Eq. (75). This result explains why filling of solenoids with soft ferromagnets with $\mu \gg \mu_0$ is so popular in the electrical engineering practice, where large self- and mutual inductances are frequently needed in systems with size and/or weight restrictions.

Now, let us use these two tools to discuss a curious (and practically important) approach to systems with ferromagnetic cores. First, let us find the magnetic flux Φ in a system with a relatively thin, closed magnetic core made of sections of (possibly, different) soft ferromagnets, with the cross-section areas A_k much smaller than the squared lengths l_k of the sections - see Fig. 15.

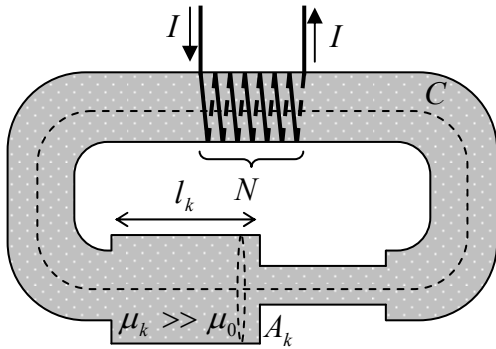


Fig. 5.15. Deriving the “magnetic Ohm law” (135).

If all $\mu_k \gg \mu_0$, virtually all field lines are confined to the interior of the core. Then, applying the macroscopic Ampère law (117) to contour C , which follows a magnetic field line inside the core (see the

dashed line in Fig. 15), we get the following approximate expression (exactly valid only in the limit $\mu_k/\mu_0, l_k^2/A_k \rightarrow \infty$):

$$\oint_C H_l dl \approx \sum_k l_k H_k = \sum_k l_k \frac{B_k}{\mu_k} = NI. \quad (5.134)$$

However, since the magnetic field lines stay in the core, the magnetic flux $\Phi_k \approx B_k A_k$ should be the same ($\equiv \Phi$) for each section, so that $B_k = \Phi/A_k$. Plugging this condition into Eq. (134), we get

Magnetic
Ohm law
and
reluctance

$$\Phi = \frac{NI}{\sum_k \mathcal{R}_k}, \quad \text{where } \mathcal{R}_k \equiv \frac{l_k}{\mu_k A_k}. \quad (5.135)$$

Note a close analogy of the first of these equations with the Ohm law for several resistors connected in series, with the magnetic flux playing the role of electric current, while the product NI , of the voltage applied to the resistor chain. This analogy is fortified by the fact that the second of Eqs. (135) is similar to the expression for resistance $R = l/\sigma A$ of a long uniform conductor, with the magnetic permeability μ playing the role of the electric conductivity σ . (In order to sound similar, but still different from resistance R , parameter \mathcal{R} is called the *reluctance*.) This is why Eq. (135) is called the *magnetic Ohm law*; it is very useful for approximate analyses of systems like ac transformers, magnetic energy storage systems, etc.

The role of the “magnetic e.m.f.” NI may be also played by a permanent-magnet section of the core. Indeed, for relatively low fields we may use the Taylor expansion of the nonlinear function $B(H)$ near $H = 0$ to write

$$B \approx \mp \mu_0 M_s + \mu_d H, \quad \mu_d \equiv \left. \frac{dB}{dH} \right|_{H=0}, \quad (5.136)$$

where M_s is the spontaneous magnetization magnitude at $H = 0$, the \mp sign corresponds to two possible directions of the magnetization, and parameter μ_d is called the *differential* (or “dynamic”) permeability. Expressing H from this relation, and using it in one of components of the sum (134), we again get a result similar to Eq. (135)

$$\Phi = \mp \frac{(NI)_{\text{ef}}}{\mathcal{R}_H + \sum_k \mathcal{R}_k}, \quad \text{with } \mathcal{R}_H \equiv \frac{l_H}{A_H \mu_d}, \quad (5.137)$$

where l_H and A_H are geometric dimensions of the hard-ferromagnet section, and product NI is replaced with its effective value

$$(NI)_{\text{ef}} = \mp \frac{\mu_0}{\mu_d} M_s l_H. \quad (5.138)$$

This result may be used for a semi-quantitative explanation of the well-known short-range forces acting between permanent magnets (or between them and soft ferromagnets) at their mechanical contact (Fig. 16).

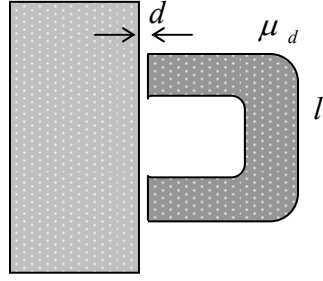


Fig. 5.16. Short-range interaction between magnets.

Indeed, considering the free-space gaps between them as sections of the core (which is approximately correct, because due to the small gap thickness d the magnetic field lines cannot stray far from the contact area), and neglecting the reluctance \mathcal{R} of the bulk material (due to its larger cross-section), we get

$$|\Phi| \propto \left(\frac{2d}{\mu_0} + \frac{l}{\mu_d} \right)^{-1}, \quad (5.139)$$

so that, according to Eq. (127), the magnetic energy of the system (disregarding the constant energy of the permanent magnetization) is

$$U \propto \left(\frac{2d}{\mu_0} + \frac{l}{\mu_d} \right) B^2 \propto \left(\frac{2d}{\mu_0} + \frac{l}{\mu_d} \right)^{-1} \propto \frac{1}{d + d_0}, \quad d_0 \equiv \frac{1}{2} \frac{\mu_0}{\mu_d} l \ll l. \quad (5.140)$$

Hence the magnet attraction force,

$$F = \left| \frac{\partial U}{\partial d} \right| \propto \frac{1}{(d + d_0)^2}, \quad (5.141)$$

behaves almost as the divergence $1/d^2$ truncated at a short distance $d_0 \ll l$. Due to that truncation, the force is finite at $d = 0$; this exactly the force you need to apply to detach two magnets.

Finally, let us discuss in brief a related effect in experiments with thin and long hard ferromagnetic samples - “needles”, like those used in magnetic compasses. Using the definition (108) of field \mathbf{H} , the Maxwell equation (29) takes the form

$$\nabla \cdot \mathbf{B} \equiv \mu_0 \nabla \cdot (\mathbf{H} + \mathbf{M}) = 0, \quad (5.142)$$

and may be rewritten as

$$\nabla \cdot \mathbf{H} = -\nabla \cdot \mathbf{M}. \quad (5.143)$$

While this relation is general, it is especially convenient in hard ferromagnets, where \mathbf{M} is virtually fixed by the saturation. Comparing this equation with Eq. (1.27) for the electrostatic field, we see that the right-hand part of Eq. (143) may be considered as a fixed source of a Coulomb-like magnetic field.

For example, let us apply Eq. (143) to a thin, long needle made of a hard ferromagnet (Fig. 17a). Inside the needle, $\mathbf{M} = \mathbf{M}_s = \text{const}$, while outside it $\mathbf{M} = 0$, so that the right-hand part of Eq. (143) is substantially different from zero only in two small areas at the needle’s ends, and on much larger distances we can use the following approximation:

$$\nabla \cdot \mathbf{H} = -q_m \delta(\mathbf{r} - \mathbf{r}_1) + q_m \delta(\mathbf{r} - \mathbf{r}_2), \quad (5.155)$$

where $\mathbf{r}_{1,2}$ are ends' positions, and $q_m \equiv M_s A$, with A being the needle's cross-section area. This equation is completely similar to Eq. (1.27) for the electric field created by two equal and opposite point charges. In particular, if two ends of two needles are held at an intermediate distance r ($A^{1/2} \ll r \ll l$, where l is the needle length, see Fig. 17b), the ends interact in accordance with the *magnetic Coulomb law*

$$F \propto \frac{q_m^2}{r^2} = \frac{M_s^2 A^2}{r^2}. \quad (5.156)$$

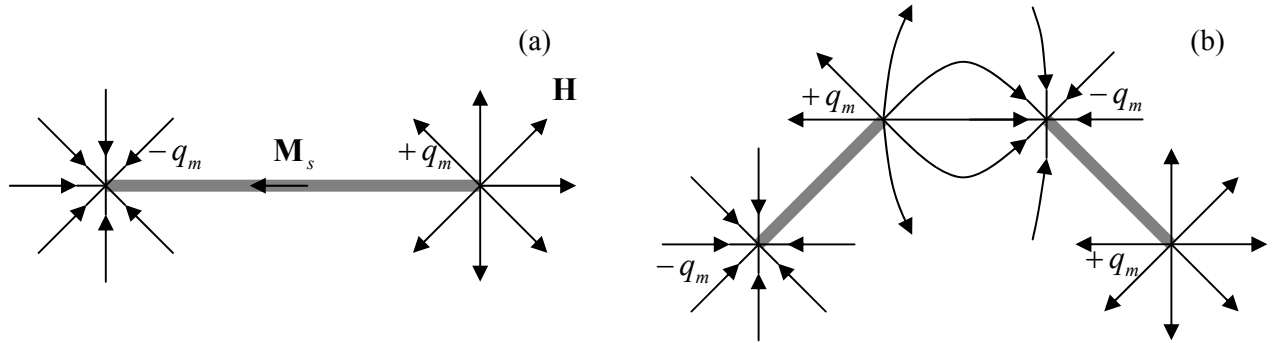
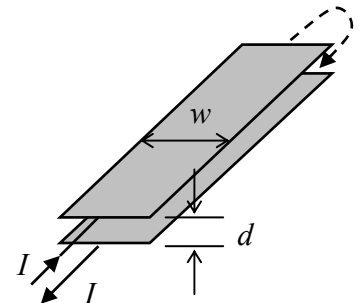


Fig. 5.17. (a) “Magnetic charges” at the ends of a thin ferromagnetic needle and (b) the result of its breaking into two parts (schematically).

The “only” (but conceptually, very significant!) difference with electrostatics is that the “magnetic charges” $\pm q_m$ cannot be fully separated. For example, if we break a magnetic needle in the middle in an attempt to bring its two ends further apart, two new “charges” appear – see Fig. 17b. There are several solid state systems where more flexible structures, similar to the magnetic needles, may be implemented. First of all, certain (“type-II”) superconductors may sustain so-called *Abrikosov vortices* – crudely, flexible tubes with field-suppressed superconductivity inside, each carrying one magnetic flux quantum $\Phi_0 = \hbar/\pi e \approx 2 \times 10^{-15}$ Wb – see Sec. 6.3. Ending on superconductor’s surface, these tubes let the magnetic field lines to spread into the surrounding space, essentially forming a magnetic monopole analog (of course, with an equal and opposite “monopole” on another end of the line). Such flux tubes are not only flexible but readily stretchable, resulting in several peculiar effects.⁵⁶ Another, recently found, examples of paired “monopoles” include *spin chains* in so-called *spin ices* – crystals with paramagnetic ions arranged into a specific (pyrochlore) lattice – such as dysprosium titanate $\text{Dy}_2\text{Ti}_2\text{O}_7$.⁵⁷

5.7. Exercise problems

5.1. Two straight, parallel, long, plane, thin strips of width d , separated by distance d , are used to form a current loop - see Fig. on the right. Calculate the magnetic field in the plane located at the middle between the planes of the strips, assuming that current I is uniformly distributed across strip width.



⁵⁶ A detailed discussion of the Abrikosov vortices may be found, for example, in Chapter 5 of M. Tinkham, *Introduction to Superconductivity*, 2nd ed., McGraw-Hill, 1996.

⁵⁷ See, e.g., L. Jaubert and P. Holdworth, *J. Phys. – Cond. Matt.* **23**, 164222 (2011) and references therein.

5.2. For the system studied in the previous problem, but now only in the limit $d \ll w$, calculate:

- (i) the distribution of the magnetic field (in the simplest possible way),
- (ii) the vector-potential of the field,
- (iii) the force (per unit length) acting on each strip, and
- (iv) the magnetic energy and self-inductance of the system (per unit length).

5.3. Calculate the magnetic field distribution near the center of the system of two similar, plane, round, coaxial wire coils, fed by equal but oppositely directed currents – see Fig. on the right.

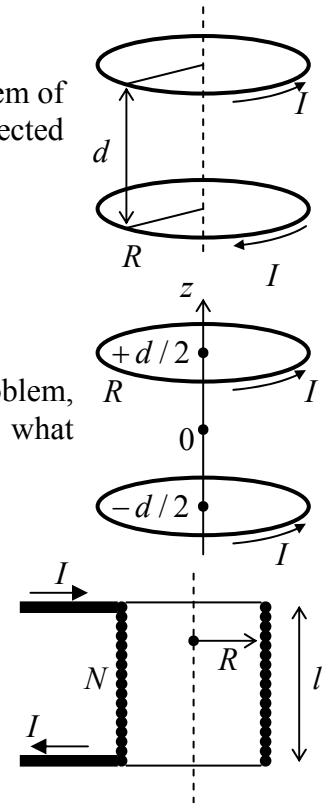
5.4. The two-coil-system, similar to that considered in the previous problem, carries equal and similarly directed currents – see Fig. on the right. Calculate what should be the ratio d/R for the second derivative $\partial^2 B_z / \partial z^2$ at $z = 0$ to vanish.⁵⁸

5.5. Calculate the magnetic field distribution along the axis of a straight solenoid (see Fig. 6a, partly reproduced on the right) with a finite length l , and round cross-section of radius R . Assume that the solenoid has many wire turns ($N \gg 1$) that are uniformly distributed along its length.

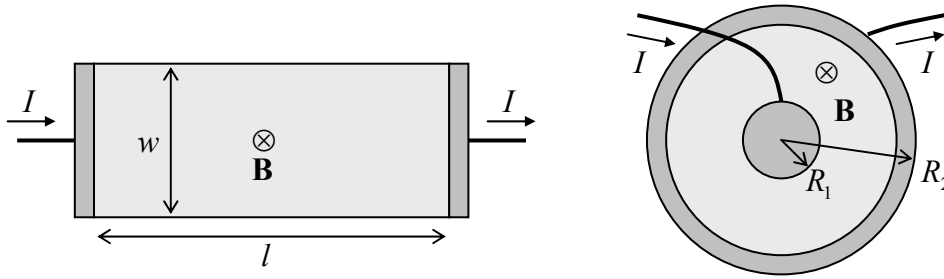
5.6. A thin spherical shell of radius R , with charge Q uniformly distributed over its surface, rotates about its axis with angular velocity ω . Calculate the distribution of the magnetic field everywhere in space.

5.7. A sphere of radius R , made of an insulating material with a uniform electric charge density ρ , rotates about its diameter with angular velocity ω . Calculate the magnetic field distribution inside the sphere and outside it.

5.8. The reader is (hopefully :-) familiar with the classical Hall effect when it takes place in the usual rectangular *Hall bar* geometry – see the left panel of the Fig. below. However, the effect takes a different form in the so-called *Corbino disk* – see the right panel below. (Dark shading shows electrodes, with no appreciable resistance.) Analyze the effect in both geometries, assuming that in both cases the conductors are thin, planar, have a constant Ohmic conductivity σ and charge carrier density n , and that the applied magnetic field \mathbf{B} is uniform and normal to conductors' planes.



⁵⁸ Such system, producing a highly uniform field near its center, is called the *Helmholtz coils*, and is broadly used in physics experiment.

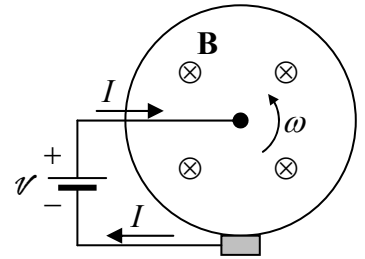


5.9.* The simplest model of the famous *homopolar motor*⁵⁹ is a thin, round conducting disk, placed into a uniform magnetic field normal to its plane, and fed by dc current flowing from disk's center to a sliding electrode ("brush") – see Fig. on the right.

(i) Express the torque, rotating the disk, via its radius R , magnetic field \mathbf{B} , and current I .

(ii) If the disk is allowed to rotate about its axis, and the motor is driven by a battery with e.m.f. \mathcal{V} , calculate its angular velocity ω , neglecting electric circuit's resistance and friction.

(iii) Now assuming that the current circuit (battery + wires + contacts + disk itself) has full resistance R , derive and solve the equation for the time evolution of ω , and analyze the solution.



5.10.* Estimate the values of magnetic susceptibility due to

- (i) orbital diamagnetism, and
- (ii) spin paramagnetism,

for a dilute medium with negligible interaction between molecular dipoles.

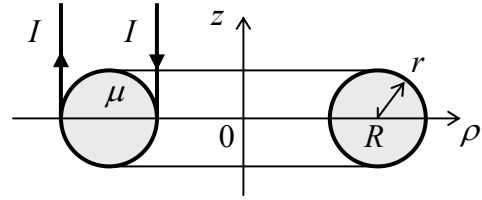
Hints: For task (i), you may use the classical model described by Eq. (114) (see Fig. 13), while for task (ii), assume the mechanism of ordering of spontaneous magnetic dipoles \mathbf{m}_0 , similar to the one sketched for electric dipoles in Fig. 12b, with the magnitude of the order of the Bohr magneton μ_B – see Eq. (96).

5.11.* Use the classical picture of the orbital ("Larmor") diamagnetism, discussed in Sec. 5.5 of the lecture notes, to calculate its (small) correction $\Delta\mathbf{B}(0)$ to the magnetic field \mathbf{B} , as felt by the atomic nucleus, modeling atomic electrons by a spherically-symmetric cloud with electric charge density $\rho(r)$. Express the result via the value $\phi(0)$ of the electrostatic potential of electrons' cloud, and use this expression for a crude numerical estimate of the relative correction, $\Delta B(0)/B$, for the hydrogen atom.

5.12. Current I is flows in a thin wire bent into a plane, round loop of radius R . Calculate the net magnetic flux through the whole plane in which the loop is located.

⁵⁹ It was invented by M. Faraday in 1821, i.e. well before his celebrated work on electromagnetic induction. The adjective "homopolar" refers to the constant "polarity" (sign) of the current; the alternative term is "unipolar".

5.13. Calculate the (self-) inductance of a toroidal solenoid (Fig. 6b) with the round cross-section of radius $r \sim R$ (see Fig. on the right), filled with a material of magnetic permeability μ , with many ($N \gg 1$, R/r) wire turns uniformly distributed along the perimeter. Check your results by analyzing the limit $r \ll R$.



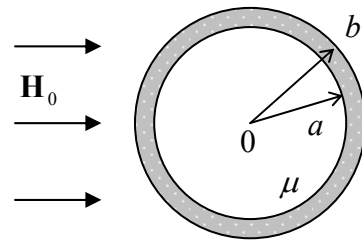
Hint : You may like to use the following table integral:⁶⁰

$$\int_0^1 \ln \frac{a + (1 - \xi^2)^{1/2}}{a - (1 - \xi^2)^{1/2}} d\xi = \pi \left[a - (a^2 - 1)^{1/2} \right], \quad \text{for } a \geq 1.$$

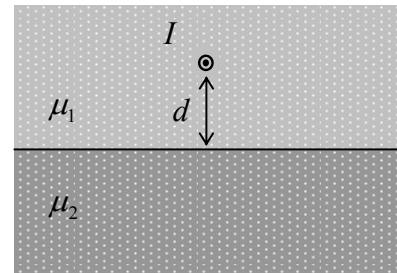
5.14. Prove that:

- (i) the self-inductance L of a current loop cannot be negative, and
- (ii) each mutual inductance coefficient $L_{kk'}$, defined by Eq. (60), cannot be larger than $(L_{kk}L_{k'k'})^{1/2}$.

5.15. A round cylindrical shell, made of a soft ferromagnet, is placed into a uniform external field \mathbf{H}_0 perpendicular to its axis - see Fig. on the right. Find the distribution of the magnetic field everywhere in the system, and discuss its efficiency as a “magnetic shield”.



5.16. A straight thin wire, carrying current I , passes parallel to the plane boundary between two uniform, linear magnetics – see Fig. on the right. Calculate the magnetic field everywhere in the system, and the force (per unit length) exerted on the wire.

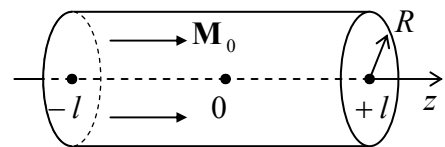


5.17. Calculate the distribution of magnetic field around a sphere made of a hard ferromagnet with a permanent, uniform magnetization $\mathbf{M} = \text{const}$.

5.18.* A limited volume V is filled with a magnetic material with magnetization $\mathbf{M}(\mathbf{r})$.

- (i) Use Eq. (5.143) to write explicit expressions for the magnetic field and its potential, induced by the magnetization.
- (ii) Recast these expressions in forms convenient when $\mathbf{M}(\mathbf{r}) = \mathbf{M}_0 = \text{const}$ inside volume V .

5.19. Use the results of the previous problem to calculate the distribution of the magnetic field along the axis of a straight



⁶⁰ See, e.g., MA (6.13).

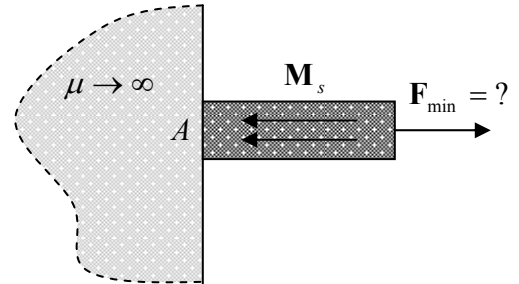
permanent magnet of length $2l$, with round cross-section of radius R , and uniform magnetization \mathbf{M}_0 parallel to the axis - see Fig. on the right.

5.20. A very broad film of thickness $2t$ is magnetized normally to its plane, with a periodic checkerboard pattern with square side a :

$$\mathbf{M}|_{|z|<t} = \mathbf{n}_z M(x, y), \quad \text{with } M(x, y) = M_0 \times \begin{cases} (+1), & \text{if } \cos \frac{\pi x}{a} \cos \frac{\pi y}{a} > 0, \\ (-1), & \text{if } \cos \frac{\pi x}{a} \cos \frac{\pi y}{a} < 0. \end{cases}$$

Calculate the magnetic field distribution in space.⁶¹

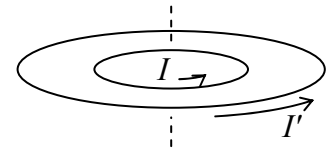
5.21. A flat end of a rod magnet, with cross-section area A , with saturated magnetization \mathbf{M}_s directed along rod's length, is let to stick to a plane surface of a large sample made of a soft ferromagnetic material with $\mu \gg \mu_0$. Calculate the force necessary to detach the rod from the surface, if it is applied strictly perpendicular to the contact surface – see Fig. on the right.



5.22.* Based on the discussion of the quadrupole electrostatic lens in Sec. 2.4 of the lecture notes, suggest permanent-magnet systems which may similarly focus particles moving close to system's axis, and carrying:

- (i) an electric charge,
- (ii) no net electric charge, but a nonvanishing spontaneous magnetic dipole moment \mathbf{m} .

5.23. A circular wire loop, carrying a fixed dc current, has been placed inside a similar but larger loop, carrying a fixed current in the same direction – see Fig. on the right. Use semi-quantitative arguments to analyze the mechanical stability of the coaxial, coplanar position of the inner loop with respect to its possible angular, axial, and lateral displacements, if the position of the outer loop is fixed.



⁶¹ This problem is of an evident relevance for the *perpendicular magnetic recording* (PMR) technology, which presently dominates the high-density digital magnetic recording, with the density already approaching 1 Tb/in².

This page is
intentionally left
blank

Chapter 6. Time-Dependent Electromagnetism

In this chapter discusses two major new effects that appear if the electric and magnetic fields are changing in time: the “electromagnetic induction” of electric field by changing magnetic field, and the reciprocal effect of “displacement currents” - the induction of magnetic field by changing electric field. These two phenomena, which make the time-dependent electric and magnetic fields inseparable, contribute to the system of four Maxwell equations, and make it valid for arbitrary electromagnetic processes. On the way, I will pause for a brief review of the electrodynamics of superconductivity, which (besides its own significance), provides a perfect platform for a discussion of the gauge invariance.

6.1. Electromagnetic induction

As Eqs. (5.36) and (5.109) show, in static situations ($\partial/\partial t = 0$) the Maxwell equations describing the electric and magnetic fields are independent, and are coupled only implicitly, via the continuity equation (4.5) relating their right-hand parts ρ and \mathbf{j} . (In statics this relation imposes a restriction only on vector \mathbf{j} .) In dynamics, when the fields change in time, the situation is different.

Historically, the first discovered explicit coupling between the electric and magnetic fields was the effect of electromagnetic induction.¹ The summary of Faraday’s numerous experiments has turned out to be very simple: if the magnetic flux, defined by Eq. (5.65),

$$\Phi \equiv \int_S \mathbf{B}_n d^2r, \quad (6.1)$$

through a surface S limited by contour C , changes in time by whatever reason (e.g., either due to a change of the magnetic field \mathbf{B} , or contour’s motion, or its deformation), it induces an additional, vortex-like electric field \mathbf{E}_{ind} , similar in its topology to the magnetic field induced by a current. The exact distribution of \mathbf{E}_{ind} in space depends on system geometry details and may be rather complex, but its integral along the contour C , called the *inductive electromotive force* (e.m.f.), obeys a very simple *Faraday induction law*:²

$$\mathcal{V}_{\text{ind}} \equiv \oint_C \mathbf{E}_{\text{ind}} \cdot d\mathbf{r} = -\frac{d\Phi}{dt}. \quad (6.2)$$

Faraday
induction
law

It is straightforward (and hence left for the reader’s exercise :-)) to show that the e.m.f. may be measured, for example, either inserting a voltmeter into a conducting loop following contour C , or by measuring current $I = \mathcal{V}_{\text{ind}}/R$ it induces in a thin wire with Ohmic resistance R , whose shape follows that contour. The minus sign in Eq. (2) corresponds to the so-called *Lenz rule*: the magnetic field of the induced Ohmic current provides a partial compensation of the change of the original Φ in time.

In order to recast Eq. (2) in a differential form, let us apply, to the above definition of \mathcal{V}_{ind} , the same Stokes theorem that was repeatedly used in Chapter 5.³ The result is

¹ The induction e.m.f. was discovered independently by J. Henry and M. Faraday, but it was a brilliant experiment series of the latter physicist, carried out in 1831, which resulted in this general formulation of the law.

² In Gaussian units, the right-hand part of this formula has the additional coefficient $1/c$.

³ If necessary, see MA Eq. (12.1) again.

$$\mathcal{V}_{\text{ind}} = \int_S (\nabla \times \mathbf{E}_{\text{ind}})_n d^2r. \quad (6.3)$$

Now combining Eqs. (1)-(3), for a contour C whose shape does not change in time (so that the integration along it is interchangeable with the time derivative),⁴ we get

$$\int_S \left(\nabla \times \mathbf{E}_{\text{ind}} + \frac{\partial \mathbf{B}}{\partial t} \right)_n d^2r = 0. \quad (6.4)$$

Since the induced electric field is additional to the field (1.33) created by electric charges, for the net field we should write $\mathbf{E} = \mathbf{E}_{\text{ind}} - \nabla\phi$. However, since curl of any gradient field is zero,⁵ $\nabla \times (\nabla\phi) = 0$, Eq. (4) is valid for the net field \mathbf{E} . Since this equation should be correct for *any* closed area S , we may conclude that

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 \quad (6.5)$$

Differential
form of the
Faraday law

at any point. This is the final (time-dependent) form of this Maxwell equation. Superficially, it may look that Eq. (5) is less general than Eq. (2); for example that it does not describe any electric field, and hence any e.m.f. in a moving loop, if field \mathbf{B} is constant in time, so that flux (1) does change in time. However, this is not true; in Chapter 9 we will see that in the reference frame moving with the loop such e.m.f. does appear.

Now let us re-formulate Eq. (5) in terms of the vector-potential. Since the induction effect does not alter the fundamental relation $\nabla \cdot \mathbf{B} = 0$, we still may present the magnetic field as prescribed by Eq. (5.27), i.e. as $\mathbf{B} = \nabla \times \mathbf{A}$. Plugging this expression into Eq. (6), we get

$$\nabla \times \left(\mathbf{E} + \frac{\partial \mathbf{A}}{\partial t} \right) = 0. \quad (6.6)$$

Hence we can use the argumentation of Sec. 1.3 (there applied to vector \mathbf{E} alone) to present the expression in parentheses as $-\nabla\phi$, so that

$$\mathbf{E} = -\frac{\partial \mathbf{A}}{\partial t} - \nabla\phi. \quad (6.7)$$

Electric
field vs.
potentials

It is tempting to interpret the first term of the right-hand part as describing the electromagnetic induction alone, and the second term representing a purely electric field induced by electric charges. However, the separation of these two terms is, to a certain extent, conditional. Indeed, let us consider the gauge transformation already mentioned in Sec. 5.2,

$$\mathbf{A} \rightarrow \mathbf{A} + \nabla\chi, \quad (6.8)$$

⁴ Let me admit that from the beginning of the course, I was carefully sweeping under the rug a very important question: in what exactly reference frame(s) all the equations of electrodynamics are valid? I promise to discuss this issue in detail later in the course (in Chapter 9), and for now would like to get away with a very short answer: all the formulas discussed so far are valid *any inertial* reference frame, as defined in classical kinematics – see, e.g., CM Chapter 1. It is crucial, however, to have fields \mathbf{E} and \mathbf{B} measured *in the same* reference frame.

⁵ See, e.g., MA Eq. (11.1).

that, as we already know, does not change the magnetic field. According to Eq. (8), in order to keep the full electric field intact (*gauge-invariant*) as well, the scalar electric potential has to be transformed simultaneously, as

$$\phi \rightarrow \phi - \frac{\partial \chi}{\partial t}, \quad (6.9)$$

leaving the choice of a time-independent addition to ϕ restricted only by the Laplace equation – since the full ϕ should satisfy the Poisson equation (1.41) with a gauge-invariant right-hand part. We will return to the discussion of gauge invariance in Sec. 3.

Now let us discuss whether Eqs. (2) or (5) describing the electromagnetic induction represent some completely new facts, on top of all the equations of electrostatics and magnetostatics, discussed in previous five chapters. The answer is *not*. To demonstrate that, let us consider a thin wire loop with current I , placed in a magnetic field (Fig. 1). According to Eq. (5.21), the magnetic force exerted by the field upon a small fragment of the wire is

$$d\mathbf{F} = I(d\mathbf{r} \times \mathbf{B}) = -I(\mathbf{B} \times d\mathbf{r}), \quad (6.10)$$

where $d\mathbf{r}$ is a small vector, tangential to loop's contour and directed along current I . Now let the wire be slightly (and slowly) deformed so that this particular fragment is displaced by a small distance $\delta\mathbf{r}$. (Let me hope that Fig. 1 makes the difference between the elementary vectors $d\mathbf{r}$ and $\delta\mathbf{r}$ absolutely clear.)

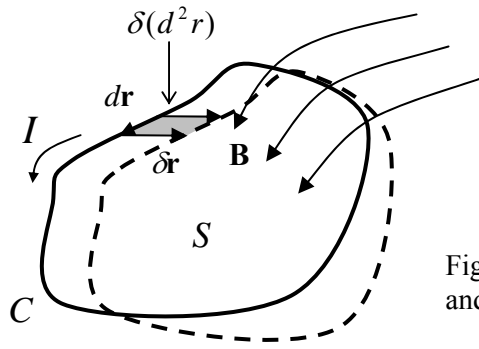


Fig. 6.1. Thin wire with current in a magnetic field, and its small deformation.

Since the wire's acceleration (if any) is negligibly small, *external* (non-magnetic) forces should balance force (10), i.e. provide an equal and opposite force. This is why the work of these external forces at the displacement $\delta\mathbf{r}$, i.e. the change of the magnetic field energy U , is,

$$\delta(dU) = -d\mathbf{F} \cdot \delta\mathbf{r} = I\delta\mathbf{r} \cdot (\mathbf{B} \times d\mathbf{r}). \quad (6.11)$$

Let us apply to this mixed product the general operand rotation rule of the vector algebra,⁶ so that vector \mathbf{B} comes out of the vector product:

$$\delta(dU) = I\mathbf{B} \cdot (d\mathbf{r} \times \delta\mathbf{r}). \quad (6.12)$$

But the magnitude of this vector product is nothing more than the area $\delta(d^2 r) \equiv \delta(dS)$ swept by the wire's fragment at the deformation (Fig. 1), while its direction is perpendicular to this elementary area dS , along the “proper” normal vector $\mathbf{n} = (d\mathbf{r}/dr) \times (\delta\mathbf{r}/\delta r)$. The scalar multiplication of \mathbf{B} by this vector is

⁶ See, e.g., MA Eq. (7.6).

equivalent to taking its normal component. Hence, integrating Eq. (12) over all the wire length, we get the following result for the total variation of the magnetic energy:

$$\delta U = I \oint_C \mathbf{B}_n \delta(d^2 r). \quad (6.13)$$

If \mathbf{B} does not change at the wire deformation, the variation sign may be moved out from the integral, and Eq. (13) yields⁷

$$\delta U = I \delta \Phi, \quad (6.14)$$

where Φ is the magnetic flux through the loop.

Now let the work $\delta \mathcal{W} = \delta U$, necessary for this energy change, to come from a generator of voltage V_{ext} , inserted somewhere in the loop. In order for the system to be in quasi-equilibrium, this voltage should counter-balance the electromagnetic induction's e.m.f. \mathcal{V}_{ind} . Work of the voltage at transfer of charge $\delta Q = I \delta t$, during elementary deformation's duration δt , is

$$\delta \mathcal{W} = V_{\text{ext}} \delta Q = -\mathcal{V}_{\text{ind}} \delta Q = -\mathcal{V}_{\text{ind}} I \delta t. \quad (6.15)$$

Comparing Eqs. (14) and (15), we arrive at the Faraday induction law (2).

Moreover, some authors *derive* Eq. (2) in this way, implying that there is no new information in the induction law at all. Note, however, that the simple derivation given above has used the assumption of magnetic field's independence on the deformation. A removal of this limitation would require using the Lorentz field transform (which will be only discussed in Chapter 9), and a very careful argumentation to exclude a faulty logic loop, because the transform itself is typically derived from Maxwell equations - including Eq. (5) that we are trying to prove. Personally I am happy that Dr. Faraday did his thorough work so early, placing the electromagnetic induction law on a firm experimental basis.

6.2. Quasistatic approximation and skin effect

As we will see later in this chapter, the interplay of the electromagnetic induction with one more time-dependent effect (the so-called *displacement currents*), enables electromagnetic waves propagating with speed $c = 1/(\epsilon_0 \mu_0)^{1/2}$ in free space, and with a comparable speed $v = 1/(\epsilon \mu)^{1/2}$ in dielectric and/or magnetic materials. For the phenomena whose spatial scale is much smaller than the wavelength $\lambda = 2\pi v/\omega$, the displacement current effects are negligible, and time-dependent phenomena may be described by using Eq. (6) together with three other macroscopic Maxwell equations in their unmodified form:⁸

$$\begin{aligned} \nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} &= 0, & \nabla \times \mathbf{H} &= \mathbf{j}, \\ \nabla \cdot \mathbf{D} &= \rho, & \nabla \cdot \mathbf{B} &= 0. \end{aligned} \quad (6.16) \quad \text{Quasistatic approximation}$$

These equations define the so-called *quasistatic approximation* of electromagnetism, and are sufficient to describe many important phenomena. Let us use them first of all for an analysis of the so-

⁷ Actually, Eq. (14) is just an integral version of Eq. (5.128).

⁸ Actually, the absence of time-dependent corrections to other Maxwell equations in the quasistatic approximation should be considered as an additional experimental fact.

called *skin effect*, the phenomenon of self-shielding of the alternating (*ac*) magnetic fields by currents flowing in a conductor.

In order to form a complete system of equations, Eqs. (16) should be augmented by material equations describing the medium. Let us take them, for a conductor, in the simplest (and simultaneously, most common) linear and isotropic form:

$$\mathbf{j} = \sigma \mathbf{E}, \quad \mathbf{B} = \mu \mathbf{H}. \quad (6.17)$$

If the conductor is uniform, i.e. coefficients σ and μ are constant inside it, the whole system of equations (16)-(17) may be reduced to a single equation. Indeed, a sequential substitution of these equations into each other yields:

$$\begin{aligned} \frac{\partial \mathbf{B}}{\partial t} &= -\nabla \times \mathbf{E} = -\frac{1}{\sigma} \nabla \times \mathbf{j} = -\frac{1}{\sigma} \nabla \times (\nabla \times \mathbf{H}) = -\frac{1}{\sigma \mu} \nabla \times (\nabla \times \mathbf{B}) = -\frac{1}{\sigma \mu} [\nabla(\nabla \cdot \mathbf{B}) - \nabla^2 \mathbf{B}] \\ &= \frac{1}{\sigma \mu} \nabla^2 \mathbf{B}. \end{aligned} \quad (6.18)$$

Thus we have arrived, without any further assumptions, at a very simple partial differential equation. Let us use it to analyze the skin effect in the simplest geometry (Fig. 2a) when an external source (which, at this point, does not need to be specified) produces, near a plane surface of a bulk conductor, a spatially-uniform *ac* magnetic field $\mathbf{H}^{(0)}(t)$ parallel to the surface.

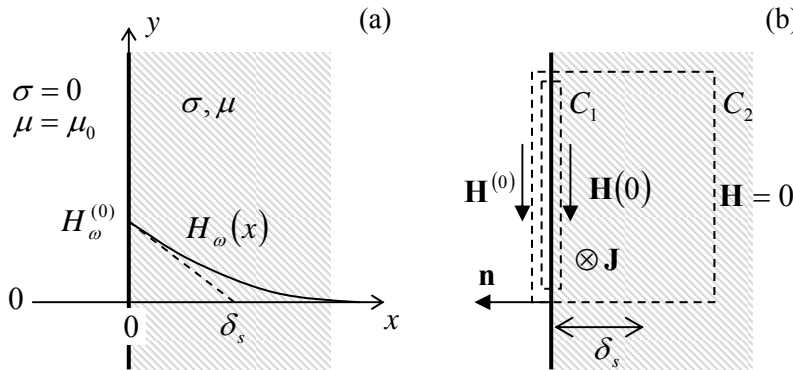


Fig. 6.2. (a) Skin effect in the simplest, planar geometry, and (b) two Ampère contours for deriving the “microscopic” (contour C_1) and the “macroscopic” (contour C_2) boundary conditions for \mathbf{H} .

Selecting the coordinate system as shown in Fig. 2, we may express this condition as

$$\mathbf{H}|_{x=0} = H^{(0)}(t) \mathbf{n}_y. \quad (6.19)$$

The translational symmetry of our simple problem within the surface plane $[y, z]$ implies that inside the conductor $\partial/\partial y = \partial/\partial z = 0$ as well, and $\mathbf{H} = H(x, t) \mathbf{n}_y$ even at $x \geq 0$, so that Eq. (18) for conductor’s interior is reduced to a differential equation for just one scalar function $H(x, t) = B(x, t)/\mu$:⁹

$$\frac{\partial H}{\partial t} = \frac{1}{\sigma \mu} \frac{\partial^2 H}{\partial x^2}, \quad \text{for } x \geq 0. \quad (6.20)$$

⁹ Due to the simple linear relation between fields \mathbf{B} and \mathbf{H} , it does not matter too much which of them is used for the solution of this problem. A slight preference is for \mathbf{H} , due to the simplicity of the boundary condition (5.118).

This equation may be further simplified by noticing that due to its linearity, we may use the linear superposition principle for the time dependence of the field,¹⁰ via expanding it, as well as the external field (19), into the Fourier series,

$$\begin{aligned} H(x, t) &= \sum_{\omega} H_{\omega}(x) e^{-i\omega t}, \quad \text{for } x \geq 0, \\ H^{(0)}(t) &= \sum_{\omega} H_{\omega}^{(0)} e^{-i\omega t}, \quad \text{for } x = 0, \end{aligned} \quad (6.21)$$

and arguing that if we know the solution for each frequency component, the whole field may be found through the elementary summation (17) of these solutions. For each single-frequency component, Eq. (21) is immediately reduced to an ordinary differential equation for the complex amplitude $H_{\omega}(x)$:

$$-i\omega H_{\omega} = \frac{1}{\sigma\mu} \frac{d^2}{dx^2} H_{\omega}. \quad (6.22)$$

From the theory of linear differential equations we know that Eq. (22) has the following general solution:

$$H_{\omega}(x) = H_+ e^{\kappa_+ x} + H_- e^{\kappa_- x}, \quad (6.23)$$

where constants κ_{\pm} are roots of the characteristic equation that may be obtained by substitution of any of these two exponents into the initial differential equation. For our particular case, the characteristic equation, following from Eq. (22), is

$$-i\omega = \frac{\kappa^2}{\sigma\mu} \quad (6.24)$$

and its roots are complex constants

$$\kappa_{\pm} = (-i\mu\omega\sigma)^{1/2} = \pm \frac{1-i}{\sqrt{2}} (\mu\omega\sigma)^{1/2}. \quad (6.25)$$

For our problem, the field cannot grow exponentially at $x \rightarrow +\infty$, so that only one of the coefficients, namely H_- corresponding to the decaying exponent, with $\text{Re } \kappa < 0$ (i.e. $\kappa = \kappa_-$), may be nonvanishing, so that $H_{\omega}(x) = H_{\omega}(0) \exp\{\kappa_- x\}$. In order to find the constant factor $H_{\omega}(0)$, we can integrate the Maxwell equation $\nabla \times \mathbf{H} = \mathbf{j}$ along a pre-surface contour – say, contour C_1 shown in Fig. 2b. The right-hand part's integral is negligible, because \mathbf{j} does not contain any “genuinely surface” currents, localized at a depth much smaller than $1/\text{Re}[-\kappa_-]$. As a result, we get the “microscopic”¹¹ boundary condition similar to Eq. (5.118) for the stationary magnetic field, $H_{\tau} = \text{const}$ at $x = 0$, we get

$$H(0, t) = H^{(0)}(t), \quad \text{i.e. } H_{\omega}(0) = H_{\omega}^{(0)}, \quad (6.26)$$

¹⁰ Another important opportunity to exploit the linearity of Eq. (6.20) (as well as any linear, homogeneous differential equation) is to use the *spatial-temporal Green's function* approach to explore the dependence of its solutions on various initial conditions. Unfortunately, because of lack of time, I have to leave such exploration for reader's exercise.

¹¹ This common name is awkward, because Eq. (26) results from macroscopic Maxwell equations (16), but is justified as the counterpart to the “macroscopic” boundary condition (30), to be discussed in a minute.

so that the final solution of the problem may be presented as

$$H_{\omega}(x) = H_{\omega}^{(0)} \exp\left\{-\frac{x}{\delta_s}\right\} \exp\left\{-i\left(\omega t - \frac{x}{\delta_s}\right)\right\}, \quad (6.27a)$$

where constant δ_s is called the *skin depth*:

Skin
depth

$$\delta_s \equiv -\frac{1}{\operatorname{Re} \kappa_-} = \left(\frac{2}{\mu \sigma \omega}\right)^{1/2}. \quad (6.27b)$$

This solution describes the *skin effect*: the penetration of the ac magnetic field of frequency ω into a conductor only to a depth of the order of δ_s . A couple of examples of the skin depth: for copper at room temperature, $\delta_s \approx 1$ cm at the ac power distribution frequency of 60 Hz, and is of the order of just 1 μm at a few GHz, i.e. at typical frequencies of cell phone signals and kitchen microwave magnetrons. For a modestly salted water, δ_s is close to 250 m at 1 Hz (with big implications for radio communications with submarines), and is of the order of 1 cm at a few GHz (explaining a nonuniform heating of a soup bowl in a microwave oven).

In order to complete the skin effect discussion, let us consider what happens with the induced ac currents¹² and the electric field at this effect. When deriving our basic equation (18), we have used, in particular, relations $\mathbf{j} = \nabla \times \mathbf{H} = \mu^{-1} \nabla \times \mathbf{B}$, and $\mathbf{E} = \mathbf{j} / \sigma$. Since a spatial differentiation of an exponent yield a similar exponents, the electric field and current density have the same spatial dependence as the magnetic field, i.e. penetrate inside the conductor by distances of the order of $\delta_s(\omega)$, but their vectors are directed perpendicularly to \mathbf{B} , while still being parallel to the conductor surface:¹³

$$\mathbf{j}_{\omega}(x) = \kappa_- H_{\omega}(x) \mathbf{n}_z, \quad \mathbf{E}_{\omega}(x) = \frac{\kappa_-}{\sigma} H_{\omega}(x) \mathbf{n}_z. \quad (6.28)$$

By the way, integrating the first of these relations with the help of Eq. (26a), we may find that the *linear density* \mathbf{J} of the surface currents (measured in A/m), is simply and fundamentally related to the applied magnetic field:

$$\mathbf{J}_{\omega} \equiv \int_0^{\infty} \mathbf{j}_{\omega}(x) dx = H_{\omega}^{(0)} \mathbf{n}_z. \quad (6.29)$$

Since this relation does not have frequency-dependent factors, we may sum it up for all frequencies and get a universal relation

$$\mathbf{J}(t) = H^{(0)}(t) \mathbf{n}_z \equiv H^{(0)}(t) (-\mathbf{n}_y \times \mathbf{n}_x) = \mathbf{H}^{(0)}(t) \times (-\mathbf{n}_x) = \mathbf{H}^{(0)}(t) \times \mathbf{n}, \quad (6.30)$$

where $\mathbf{n} = -\mathbf{n}_x$ is the outer normal to the surface – see Fig. 2b. This simple relation (whose last form is independent of the reference frame choice) is not occasional. Indeed, Eq. (30) may be readily obtained from the Ampère law (5.37) applied to a contour drawn around a fragment of the surface, but extending under it much deeper than the skin depth – see contour C_2 in Fig. 2b, regardless of the exact law of the

¹² They are frequently called *eddy currents*, because of the loop form of their lines. (In the 1D geometry explored above these loops are implicit, closing at infinity.)

¹³ Notice that vectors \mathbf{j} and \mathbf{E} are parallel, and have the same time dependence. This means that the time average of the power dissipation $\mathbf{j} \cdot \mathbf{E}$ is finite. We will return to its discussion later in this chapter.

field penetration. Relation (30) is frequently called the “macroscopic” boundary condition for the magnetic field near conductor’s surface, to distinguish it from the “microscopic” boundary condition (26).

For the skin effect, the fundamental relation between the surface current density and the external magnetic field means that the effect implementation does not require a dedicated ac magnetic field source. For example, it takes place in any wire that carries ac current, and leads to current concentration in a surface sheet of thickness $\sim \delta_s$. (Of course the quantitative analysis of this problem in a wire with an arbitrary cross-section may be technically complicated, because it requires to solve Eq. (18) for a 2D geometry; even for the round cross-section, the solution involves the Bessel functions.) In this case, the ac magnetic field outside the conductor, that still obeys Eq. (30), is better understood as the effect, rather than the reason, of the ac current flow.

Finally, the reader should mind the validity limits of these results – besides the universal Eq. (30). First, in order for the quasistatic approximation to be valid, frequency ω should not be too high, so that the skin depth (27) remains much *smaller* than the corresponding wavelength,

$$\lambda = \frac{2\pi v}{\omega} = \left(\frac{4\pi^2}{\epsilon\mu\omega^2} \right)^{1/2}, \quad (6.31)$$

which decreases with ω faster than δ_s (27b). Note that the crossover frequency (at which $\delta_s = \lambda$),

$$\omega_r = \frac{\sigma}{\epsilon} = \frac{\sigma}{\epsilon_r \epsilon_0}, \quad (6.32)$$

is nothing else than the reciprocal charge relaxation time (4.10). As was discussed in Sec. 4.2, for good metals this frequency is extremely high (about 10^{18} s^{-1}).

A more practical upper limit on ω is that the skin depth δ_s should stay much *larger* than the mean free path l of charge carriers.¹⁴ Beyond this point, a *non-local* relation between vectors $\mathbf{j}(\mathbf{r})$ and $\mathbf{E}(\mathbf{r})$ becomes essential. Both theory and experiment show that at $\delta_s < l$, the skin effect still persists, but acquires a slightly different frequency dependence, $\delta_s \propto \omega^{-1/3}$. Such *anomalous skin effect* has useful applications, for example, for experimental measurements of the Fermi surface in metals.¹⁵

6.3. Electrodynamics of superconductivity and gauge invariance

The effect of superconductivity¹⁶ takes place when temperature T is reduced below a certain *critical temperature* (T_c), specific for each material. For most metallic superconductors, T_c is of the order of typically a few kelvins, though several exotic compounds (the so-called *high-temperature superconductors*) with T_c above 100 K have been found since 1987. The most notable property of superconductors is the absence, at $T < T_c$, of measurable resistance to not very high dc currents.

¹⁴ A brief discussion of the mean free path may be found, for example, in SM Chapter 6. In very clean metals at low temperatures, δ_s may approach l at frequencies as low as $\sim 1 \text{ GHz}$, though at room temperature the crossover from the normal to the anomalous skin effect takes place at $\sim 100 \text{ GHz}$.

¹⁵ See, e.g., A. Abrikosov, *Introduction to the Theory of Normal Metals*, Academic Press, 1972.

¹⁶ Discovered experimentally in 1911 by H. Kamerlingh Onnes.

However, electromagnetic properties of superconductors cannot be described by just taking $\sigma = \infty$ in our previous results. Indeed, for this case, Eq. (27b) would give $\delta_s = 0$, i.e., no ac magnetic field penetration at all, while for the dc field we would have the uncertainty $\sigma\omega = ?$ Experiment shows something substantially different: weak magnetic fields do penetrate into superconductors by a material-specific *London penetration depth* $\delta_L \sim 10^{-7}$ - 10^{-6} m,¹⁷ which is virtually frequency-independent until the skin depth δ_s , measured in the same material in its “normal” state, i.e. the absence of superconductivity, becomes less than δ_L . (This crossover happens typically at frequencies $\sim 10^{13}$ s⁻¹.) The smallness of δ_L means that the magnetic field is pushed out of macroscopic samples at their transition into the superconducting state.

This *Meissner-Ochsenfeld effect*, discovered experimentally in 1933,¹⁸ may be partly understood using the following classical reasoning. When we discussed the physics of conductivity in Sec. 4.2, we implied that the current (and electric field) frequency ω is either zero or sufficiently low. In the classical Drude reasoning (see Sec. 4.2), this is acceptable while $\omega\tau \ll 1$, where τ is the effective carrier scattering time participating in Eqs. (4.12)-(4.13). If this condition is not satisfied, we should take into account the charge carrier inertia; moreover, in the opposite limit $\omega\tau \gg 1$ we may neglect the scattering at all. Classically, we can describe the charge carriers in such a “perfect conductor” as particles that are accelerated by the electric field in accordance with the 2nd Newton law (4.11) at all times,

$$\dot{\mathbf{v}} = \frac{1}{m} \mathbf{F} = \frac{q}{m} \mathbf{E}, \quad (6.33)$$

so that the current density $\mathbf{j} = qn\mathbf{v}$ they create changes in time as

$$\dot{\mathbf{j}} = \frac{q^2 n}{m} \mathbf{E}. \quad (6.34)$$

In terms of the Fourier amplitudes (see the previous section), this means

$$-i\omega \mathbf{j}_\omega = \frac{q^2 n}{m} \mathbf{E}_\omega. \quad (6.35)$$

Comparing this formula with the relation $\mathbf{j}_\omega = \sigma \mathbf{E}_\omega$ implied in the last section, we see that we can use all its results with the following replacement:

$$\sigma \rightarrow i \frac{q^2 n}{m\omega}. \quad (6.36)$$

This change replaces the characteristic equation (24) with

$$-i\omega = \frac{\kappa^2 m \omega}{iq^2 n \mu}, \quad (6.37)$$

i.e. replaces the skin effect with the field penetration by the following frequency-independent depth:

¹⁷ Named to acknowledge the pioneering theoretical work of brothers F. and H. London – see below.

¹⁸ It is hardly fair to shorten the name to just the “Meissner effect”, as it is frequently done, because of the reportedly crucial contribution made by R. Ochsenfeld, then W. Meissner’s student, into the discovery.

$$\delta = \left(\frac{m}{\mu q^2 n} \right)^{1/2}. \quad (6.38)$$

Superficially, this means that the field decay into the superconductor does not depend on frequency:

$$H(x, t) = H(0, t) \exp \left\{ -\frac{x}{\delta} \right\}, \quad (6.39)$$

explaining the Meissner-Ochsenfeld effect.

However, there are two problems with this result. First, for the parameters typical for good metals ($q = -e$, $n \sim 10^{29} \text{ m}^{-3}$, $m \sim m_e$, $\mu \approx \mu_0$), Eq. (38) gives $\delta \sim 10^{-8} \text{ m}$, a factor of 10 - 10^2 lower than the typical experimental values of δ_L . Experiment also shows that the penetration depth diverges at $T \rightarrow T_c$, which is not predicted by Eq. (38). Another, much more fundamental problem with Eq. (38) is that it has been derived for $\omega\tau \gg 1$. Even if we assume that somehow there are no collisions at all, i.e. $\tau = \infty$, at $\omega \rightarrow 0$ both parts of the characteristic equation (37) vanish, and we cannot make any conclusion about k . This is not just a mathematical artifact we could ignore. For example, let us place a non-magnetic metal at $T > T_c$ into a static external magnetic field. The field will completely penetrate into the sample. Now let us cool it. As soon as the temperature drops below T_c , our calculations become valid, forbidding the penetration into the superconductor of any *change* of the field, so that the initial field would be “frozen” inside the sample. The experiment shows something completely different: as T is lowered below T_c , the initial field is being pushed out of the sample.

The resolution of these contradictions has been provided by quantum mechanics. As was explained in 1957 in a seminal work by J. Bardeen, L. Cooper, and J. Schrieffer (commonly referred to the *BSC theory*), superconductivity is due to the correlated motion of electron pairs, with opposite spins and nearly opposite momenta. Such *Cooper pairs*, each with the electric charge $q = -2e$ and zero spin, may form only in a narrow energy layer near the Fermi surface, of certain thickness $\Delta(T)$. Parameter $\Delta(T)$, which may be also considered as the binding energy of the pair, tends to zero at $T \rightarrow T_c$, while at $T \ll T_c$ it has a virtually constant value $\Delta(0) \approx 3.5 k_B T_c$, of the order of a few meV for most superconductors. This fact readily explains the relatively low spatial density of the Cooper pairs: $n_p(T) \sim n\Delta(T)/\varepsilon_F \sim 10^{26} \text{ m}^{-3}$. With the correction $n \rightarrow n_p$, our Eq. (38) for the penetration depth becomes

$$\delta \rightarrow \delta_L = \left(\frac{m}{\mu q^2 n_p(T)} \right)^{1/2}. \quad (6.40) \quad \text{London penetration depth}$$

This expression diverges at $T \rightarrow T_c$, and generally fits the experimental data reasonably well, at least for the so-called “clean” superconductors (with the mean free path $l \equiv v\tau$ much longer than the Cooper pair size ξ - see below).

The smallness of the coupling energy $\Delta(T)$ is also a key factor in the explanation of the Meissner-Ochsenfeld effect, as well as several *macroscopic quantum phenomena* in superconductors. Because of Heisenberg’s quantum uncertainty relation $\delta r \delta p \sim \hbar$, the Cooper-pair size (the so-called *coherence length*) is relatively large: $\xi \sim \delta r \sim \hbar / \delta p \sim \hbar v_F / \Delta(T) \sim 10^{-6} \text{ m}$. As a result, $n_p \xi^3 \gg 1$, meaning that Cooper pairs are strongly overlapped in space. Now, due to their integer spin, Cooper pairs behave like bosons, which means in particular that at low temperature they exhibit the so-called *Bose-Einstein*

condensation onto the same energy level.¹⁹ This means that the frequency $\omega = E/\hbar$ of the time evolution of each pair's wavefunction $\Psi = \psi \exp\{-i\omega t\}$ is the same, i.e. that the phases ϕ of the wavefunctions, defined by equation

$$\psi = |\psi| e^{i\phi}, \quad (6.41)$$

become equal, so that the current is carried not by individual Cooper pairs but rather their *Bose-Einstein condensate* described by a single wavefunction. Due to this coherence, the quantum effects (which are, in usual Fermi-liquids of single electrons, masked by the statistical spread of phases ϕ), become very explicit – “macroscopic”.

To illustrate this, let us write the well-known quantum-mechanical formula for the probability current of a free, nonrelativistic particle,²⁰

$$\mathbf{j}_p = \frac{i\hbar}{2m} (\psi \nabla \psi^* - \text{c.c.}) = \frac{1}{2m} [\psi^* (-i\hbar \nabla) \psi - \text{c.c.}]. \quad (6.42)$$

Now let me borrow one result that will be proved later in the course (in Sec. 9.7) when we discuss the analytical mechanics of a charged particle moving in an electromagnetic field. Namely, in order to account for the magnetic field effects, particle's *kinetic* momentum \mathbf{p} , equal to $m\mathbf{v}$ (where $\mathbf{v} \equiv d\mathbf{r}/dt$ is particle's velocity) has to be distinguished from its *canonical* momentum,²¹

$$\mathbf{P} \equiv \mathbf{p} + q\mathbf{A}. \quad (6.43)$$

where \mathbf{A} is the vector-potential of the field – see Eq. (5.27). In contrast with Cartesian components $p_j = mu_j$ of momentum \mathbf{p} , the canonical momentum components are the generalized momenta corresponding to components r_j of the radius-vector \mathbf{r} , considered as generalized coordinates of the particle: $P_j = \partial\mathcal{L}/\partial v_j$, where \mathcal{L} is the particle's Lagrangian function. According to the general rules of transfer from classical to quantum mechanics,²² it is vector \mathbf{P} whose operator (in the Schrödinger picture) equals $-i\hbar\nabla$, so that the operator of kinetic momentum $\mathbf{p} = \mathbf{P} - q\mathbf{A}$ is equal to $-i\hbar\nabla - q\mathbf{A}$. Hence, in order to account for the magnetic field effects, we should make the following replacement:

$$-i\hbar\nabla \rightarrow -i\hbar\nabla - q\mathbf{A}. \quad (6.44)$$

In particular, Eq. (42) has to be replaced with

$$\mathbf{j}_p = \frac{1}{2m} [\psi^* (-i\hbar\nabla - q\mathbf{A}) \psi - \text{c.c.}]. \quad (6.45)$$

This expression becomes more transparent if we take the wavefunction in form (41):

¹⁹ A qualitative discussion of the Bose-Einstein condensation of bosons may be found in SM Sec. 3.4, though the full theory of superconductivity is more complex, because it describes the condensation taking place *simultaneously* with the formation of effective bosons (Cooper pairs). For a more detailed coverage of physics of superconductors, the reader may be referred, for example, to the already cited monograph by M. Tinkham, *Introduction to Superconductivity*, 2nd ed., McGraw-Hill, 1996.

²⁰ See, e.g., QM Sec. 1.4, in particular Eq. (1.47).

²¹ I am sorry to use traditional notations \mathbf{p} and \mathbf{P} for the momenta – the same symbols which were used for the electric dipole moment and polarization in Chapter 3. I hope there will be no confusion, because the latter notions are not used in this section.

²² See, e.g., CM Sec. 10.1, in particular Eq. (10.26).

$$\mathbf{j}_p = \frac{\hbar}{m} |\psi|^2 \left(\nabla \varphi - \frac{q}{\hbar} \mathbf{A} \right). \quad (6.46)$$

This relation means, in particular, that in order to keep \mathbf{j} invariant, the gauge transformation (8)-(9) has to be accompanied by a simultaneous transformation of the wavefunction phase:

$$\varphi \rightarrow \varphi + \frac{q}{\hbar} \chi. \quad (6.47)$$

It is fascinating that the quantum-mechanical wavefunction (more exactly, its phase) is *not* gauge-invariant – meaning that you may change it in your mind – at will! Again, this does not change any observable (such as \mathbf{j} or the probability density $\psi\psi^*$), i.e. any experimental results.

For the *electric* current density of the whole superconducting condensate, Eq. (46) yields

$$\mathbf{j} = \frac{\hbar q n_p(T)}{m} \left(\nabla \varphi - \frac{q}{\hbar} \mathbf{A} \right). \quad (6.48) \quad \text{Supercurrent density}$$

This equation shows that this *supercurrent* may be induced by dc magnetic field alone and does not require any electric field. Indeed, for the simplest, 1D geometry shown in Fig. 2a, $\mathbf{j}(\mathbf{r}) = j(x) \mathbf{n}_z$, $\mathbf{A}(\mathbf{r}) = A(x) \mathbf{n}_z$, and $\partial/\partial z = 0$, so that the Coulomb gauge condition (5.48) is satisfied for any choice of the gauge function $\chi(x)$, and for the sake of simplicity we can choose it to provide $\varphi(\mathbf{r}) \equiv \text{const}$,²³ so that

$$\mathbf{j} = -\frac{q^2 n_p(T)}{m} \mathbf{A}. \quad (6.49)$$

This is the so-called *London equation*, proposed (in a different form) by brothers F. and H. London in 1935 for a phenomenological description of the Meissner-Ochsenfeld effect. Combining it with Eq. (5.47), generalized for an arbitrary uniform media by the replacement $\mu_0 \rightarrow \mu$, we get

$$\nabla^2 \mathbf{A} = \frac{\mu q^2 n_p(T)}{m} \mathbf{A}. \quad (6.50)$$

This simple differential equation, similar in structure to Eq. (18), has a similar exponential solution,

$$A(x) = A(0) \exp\left\{-\frac{x}{\delta_L}\right\}, \quad B(x) = B(0) \exp\left\{-\frac{x}{\delta_L}\right\}, \quad j(x) = j(0) \exp\left\{-\frac{x}{\delta_L}\right\}, \quad (6.51)$$

that shows that the magnetic field and supercurrent penetrate into a superconductor only by the London's penetration depth δ_L , given by Eq. (40), regardless of frequency.²⁴ By the way, integrating the last result through the penetration layer, and using Eqs. (34), (43) and the vector-potential definition, $\mathbf{B} = \nabla \times \mathbf{A}$ (for our geometry, giving $B(x) = dA(x)/dx = -\delta_L A(x)$) we may check that the linear density \mathbf{J} of the surface supercurrent still satisfies the universal relation (30).

²³ This is the so-called *London gauge* which, for our geometry, is also the Coulomb gauge.

²⁴ Since not all electrons of a superconductor form Cooper pairs, at any frequency $\omega \neq 0$ they provide Joule losses which are not described by Eq. (48). These losses become very substantial when frequency ω becomes so high that the skin-effect length δ_s of the material (as measured with superconductivity suppressed, say by high magnetic field) becomes less than δ_L . For typical metallic superconductors, this happens at frequencies of a few hundred GHz, so that even for microwaves, Eq. (51) gives a fairly good description of the field penetration.

Let me hope that the physical intuition of the reader enables him or her to make the following semi-quantitative generalization of the quantitative solution (51) to superconductor sample of arbitrary shape: \mathbf{B} and \mathbf{j} may only penetrate into the sample by distances of the order of $\delta_L(0)$. In particular, for samples much larger than δ_L , the London theory gives the following “macroscopic” description of the Meissner-Ochsenfeld effect: $\mathbf{j} = 0$ and $\mathbf{B} = 0$ everywhere inside a superconductor. In this coarse description, the bulk superconductor sample behaves as an ideal diamagnet, with $\mu = 0$.²⁵ In particular, we can use this analogy and the first of Eqs. (5.125) to immediately obtain the magnetic field distribution outside a superconducting sphere:

$$\mathbf{B} = \mu_0 \mathbf{H} = -\mu_0 \nabla \phi_m, \quad \phi_m = H_0 \left(-r - \frac{R^3}{2r^2} \right) \cos \theta. \quad (6.52)$$

Figure 3 shows the corresponding surfaces of equal potential ϕ_m . It is evident that the magnetic field lines (normal to the equipotential surfaces) bend to become parallel to the superconductor’s surface. By the way, this pattern illustrates the answer to the question that might arise at making assumption (19): what happens to superconductors in a *normal* magnetic field? The answer is: the field is deformed outside the superconductor to provide $B_n = 0$ at the surface - otherwise, due to the continuity of B_n , the magnetic field would penetrate the superconductor, which is impossible. Of course this answer should be taken with a grain of salt: strong magnetic fields *do* penetrate into superconductors, destroying superconductivity (completely or partly), thus violating the Meissner-Ochsenfeld effect. Such a penetration by itself features several interesting electrodynamic effects, for whose discussion we unfortunately do not have time.²⁶

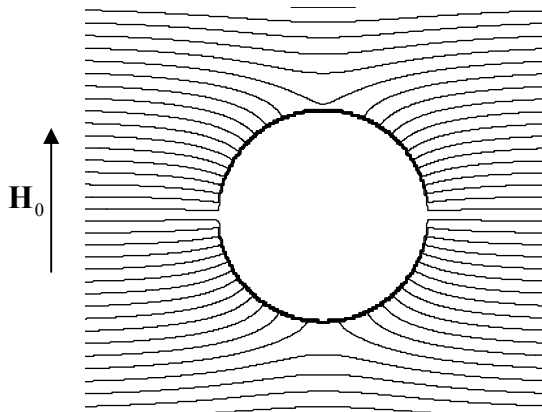


Fig. 6.3. Surfaces of constant scalar potential ϕ_m of magnetic field around a superconducting sphere of radius $R \gg \delta_L$, placed into a weak uniform, vertical magnetic field.

6.4. Electrodynamics of macroscopic quantum phenomena

We have seen that for the ac magnetic field penetration, the quantum theory of superconductivity gives essentially the same result as the classical theory of a perfect conductor – cf. Eqs. (39) and (51) – with the “only” conceptual exception that the former theory extends the effect to dc fields. However, the

²⁵ Of course, this analogy sweeps under the rug the real physics of the Meissner-Ochsenfeld effect. In particular, in superconductors the role of the surface “magnetization currents” with effective density $\mathbf{j}_{\text{ef}} = \nabla \times \mathbf{M}$ (see Fig. 5.11 and its discussion) is played by the real, persistent surface supercurrents (48).

²⁶ The interested reader may be referred, e.g., to Chapter 5 of M. Tinkham’s monograph cited above.

quantum theory of superconductors is much more rich. For example, let us use Eq. (48) to derive the fascinating effect of *magnetic flux quantization*. Consider a closed ring made of a superconducting “wire” with a cross-section much larger than δ_L^2 (Fig. 4a).

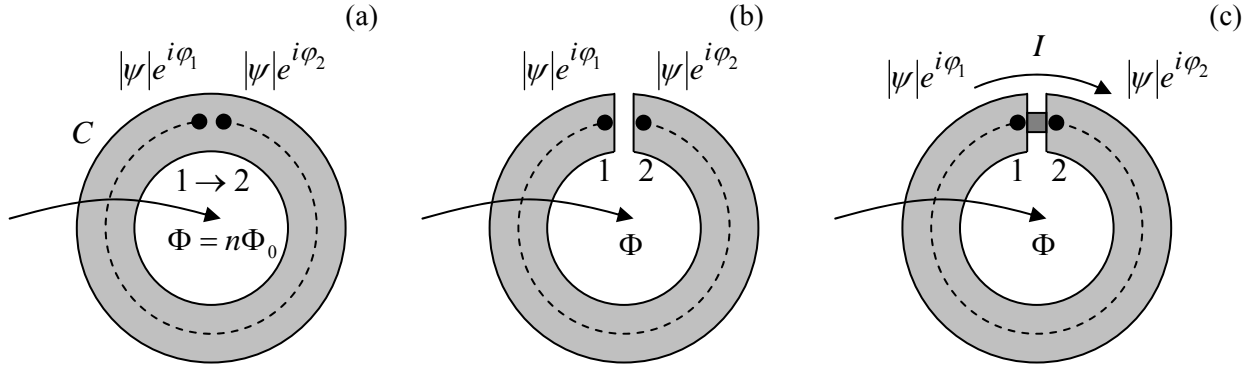


Fig. 6.4. (a) Closed, flux-quantizing superconducting ring, (b) a ring cut with a narrow slit, and (c) a Superconducting QUantum Interference Device (SQUID).

From the last section’s analysis, we know that deep inside the wire the supercurrent is exponentially small. Integrating Eq. (48) along any closed contour C that does not approach the surface closer than a few δ_L at any point, we get

$$\oint_C \nabla \varphi \cdot d\mathbf{r} - \frac{q}{\hbar} \oint_C \mathbf{A} \cdot d\mathbf{r} = 0. \quad (6.53)$$

The first integral, i.e. the difference of φ in the initial and final points, has to be equal to either zero or an integer number of 2π , because the change $\varphi \rightarrow \varphi + 2\pi m$ does not change condensate’s wavefunction:

$$\psi' = |\psi|e^{i(\varphi+2\pi m)} = |\psi|e^{i\varphi} = \psi. \quad (6.54)$$

On the other hand, the second integral in Eq. (53) is just the magnetic flux Φ (1) through the contour - and, due to the Meissner-Ochsenfeld effect, through the superconducting ring as a whole. As a result, we get

$$\Phi = n\Phi_0, \quad \Phi_0 \equiv \frac{2\pi\hbar}{q} = \frac{h}{q}, \quad n = 0, \pm 1, \pm 2, \dots, \quad (6.55)$$

Magnetic
flux
quantization

i.e. the magnetic flux can only take values multiple of the *flux quantum* Φ_0 . This effect, predicted in 1950 by the same Fritz London (who expected q to be equal to the electron charge $-e$), was confirmed experimentally in 1961,²⁷ with $|q| = 2e$ (so that in superconductors $\Phi_0 = h/2e \approx 2.07 \times 10^{-15}$ Wb). Historically, this observation gave a decisive support to the BSC theory of the Cooper pairs as the basis of superconductivity, which had been put forward just 4 years before.²⁸

²⁷ Independently and virtually simultaneously by two groups: B. Deaver and W. Fairbank, and R. Doll and M. Näbauer, so that their reports were published back-to-back in *Phys. Rev. Lett.*

²⁸ Actually, the ring is not entirely necessary. In 1957, A. Abrikosov used the Ginzburg-Landau equations (see below) to explain the counter-intuitive behavior of the so-called *type-II superconductors*, known experimentally as the *Shubnikov phase* since the 1930s. He showed that high magnetic field may penetrate into such

The flux quantization is just one of the so-called *macroscopic quantum effects* in superconductivity. Consider, for example, a superconducting ring interrupted with a very narrow slit (Fig. 4b). Integrating Eq. (48) along the current-free path from point 1 to point 2, along the dashed line in Fig. 4 (again, deeper than $\delta_L(T)$ from the surface), we get

$$0 = \int_1^2 \left(\nabla \varphi - \frac{q}{\hbar} \mathbf{A} \right) \cdot d\mathbf{r} = \varphi_2 - \varphi_1 - \frac{q}{\hbar} \Phi. \quad (6.56)$$

Using the flux quantum definition (55), this result may be rewritten as

Josephson
phase
difference

$$\varphi \equiv \varphi_1 - \varphi_2 = \frac{2\pi}{\Phi_0} \Phi, \quad (6.57)$$

where φ is called the *Josephson phase difference*. In contrast to each of the phases $\varphi_{1,2}$, their difference φ is gauge-invariant, because it is directly related to the gauge-invariant magnetic flux.

Can this φ be measured? Yes, using the *Josephson effect*.²⁹ In order to understand his prediction, let us take two (for the argument simplicity, similar) superconductors, connected with some sort of *weak link*, for example a tunnel barrier or a short normal-metal bridge, through that a small Cooper pair current can flow. (Such system of two coupled superconductors is now called a *Josephson junction*.) Let us think what this supercurrent I may be a function of. For that, the reverse thinking is helpful: let us imagine we can change current from outside; what parameter of the superconducting condensate can it affect?

If the current is weak, it cannot perturb the superconducting condensate density, proportional to $|\psi|^2$; hence it may only change the Cooper condensate phases $\varphi_{1,2}$. However, according to Eq. (41), the phases are not gauge-invariant, while the current should be, hence I may affect – or should be a function of – the *phase difference* φ defined by Eq. (57). Moreover, just has already been argued during the flux quantization discussion, a change of any of $\varphi_{1,2}$ (and hence of φ) by 2π or any of its multiples should not change the current. In addition, if the wavefunction is the same in both superconductors ($\varphi = 0$), supercurrent should vanish due to the system symmetry. Hence function $I(\varphi)$ should satisfy conditions

$$I(0) = 0, \quad I(\varphi + 2\pi) = I(\varphi). \quad (6.58)$$

With this understanding, we should not be terribly surprised by the following Josephson's result that for the weak link provided by weak tunneling,³⁰

Josephson
supercurrent

$$I(\varphi) = I_c \sin \varphi, \quad (6.59)$$

superconductors, whose coherence length ξ is smaller than the London's penetration depth $\delta_L(T)$, in the form of self-formed tubes surrounded by vortex-shaped supercurrents – the so-called *Abrikosov vortices*, with the superconductivity suppressed near the middle of each tube. This suppression makes each flux tube topologically equivalent to a superconducting ring, with the magnetic flux through it equal to one flux quantum, and its ends being magnetically similar to magnetic monopoles – see Sec. 5.6 above.

²⁹ It was predicted in 1961 by B. Josephson (then a PhD student!), and observed experimentally by several groups soon after that.

³⁰ For some other types of weak links, function $I(\varphi)$ may deviate from the sine form (59) rather considerably, still satisfying the general requirements (58).

where constant I_c , which depends on of the strength of the weak link and temperature, is called the *critical current*.

Let me show how such expression may be derived, for a narrow and short weak link made of a normal metal or a superconductor.³¹ Microscopic theory of superconductivity shows that, within certain limits, the Bose-Einstein condensate of Cooper pairs may be described by the following *nonlinear Schrödinger equation*³²

$$\frac{1}{2m}(-i\hbar\nabla - q\mathbf{A})^2\psi + U(r)\psi = \varepsilon\psi + \psi \times (\text{a nonlinear function of } |\psi|^2). \quad (6.60)$$

The first three terms of this equation are similar to those of the usual Schrödinger equation (which conserves the number of particles), while the nonlinear function in the last term describes the formation and dissolution of Cooper pairs, and in particular gives the equilibrium value of n_s as a function of temperature. Now let the weak link size scale a be much smaller than both the Cooper pair size ξ and the London's penetration depth δ_L . The first of these relations ($a \ll \xi$) makes the first term in Eq. (60), that scales as $1/a^2$, much larger than all other terms, while the latter relation ($a \ll \delta_L$) allows one to neglect magnetic field effects, and hence drop term $(-q\mathbf{A})$ from the parenthesis in Eq. (60), reducing it to just our familiar Laplace equation for the wavefunction:

$$\nabla^2\psi = 0. \quad (6.61)$$

Since the weak coupling cannot change $|\psi|$ in bulk superconducting electrodes, Eq. (61) may be solved with the following simple boundary conditions:

$$\psi(r) \rightarrow \begin{cases} |\psi|e^{i\varphi_1}, & \text{for } \mathbf{r} \rightarrow \mathbf{r}_1, \\ |\psi|e^{i\varphi_2}, & \text{for } \mathbf{r} \rightarrow \mathbf{r}_2, \end{cases} \quad (6.62)$$

where \mathbf{r}_1 and \mathbf{r}_2 are some points well inside the corresponding superconductors, i.e. at distances much larger than a from the weak link center. It is straightforward to verify that the solution of this boundary problem for *complex* function ψ may be expressed as follows,

$$\psi(\mathbf{r}) = |\psi|e^{i\varphi_1}f(\mathbf{r}) + |\psi|e^{i\varphi_2}(1 - f(\mathbf{r})), \quad (6.63)$$

via the *real* function $f(\mathbf{r})$ that satisfies the Laplace equation and the following boundary conditions:

$$f(\mathbf{r}) \rightarrow \begin{cases} 1, & \text{for } \mathbf{r} \rightarrow \mathbf{r}_1, \\ 0, & \text{for } \mathbf{r} \rightarrow \mathbf{r}_2. \end{cases} \quad (6.64)$$

Function $f(\mathbf{r})$ depends on the weak link geometry and may be rather complicated, but we do not

³¹ This derivation belongs to L. Aslamazov and A. Larkin, *JETP Lett.* **9**, 87 (1969). If the reader is not interested in this topic, he or she may safely skip it, jumping directly to the text following Eq. (68).

³² At $T \rightarrow T_c$, where $n_s \rightarrow 0$, the Taylor expansion of the nonlinear function in Eq. (60) may be limited to just one term proportional to $|\psi|^2 \propto n_s$. In this limit, Eq. (60) is called the *Ginzburg-Landau* equation – see SM (4.58). Derived by V. Ginzburg and L. Landau in 1950 from phenomenological arguments (see, e.g., SM Sec. 4.3), i.e. before the advent of the BSC theory, this simple equation, solved together with Eq. (48) and the Maxwell equations, may describe a very broad range of macroscopic quantum effects including the Abrikosov vortices, critical fields and currents, etc. – see, e.g., M. Tinkham's monograph cited above.

need to know it to get the most important result. Indeed, plugging this solution into Eq. (48) (with term $-q\mathbf{A}$ ignored as being negligibly small), we get

$$\mathbf{j}_p = -\frac{\hbar}{m}|\psi|^2 \nabla f \sin \varphi, \quad \text{so that } \mathbf{j} = -\frac{\hbar q n_p(T)}{m} \nabla f \sin \varphi. \quad (6.65)$$

Integrating this relation over any cross-section S of the weak link, we arrive at Josephson's result (59), with the following critical current:

$$I_c = -\frac{\hbar q n_p(T)}{m} \int_S (\nabla f)_n d^2 r. \quad (6.66)$$

This expression may be readily evaluated via the resistance of the same weak link in the “normal” (non-superconducting) state, say at $T > T_c$. Indeed, as we know from Sec. 4.3, the distribution of the electrostatic potential ϕ at normal conduction also obeys the Laplace equation, with boundary conditions that may be taken in the form

$$\phi(\mathbf{r}) \rightarrow \begin{cases} V, & \text{for } \mathbf{r} \rightarrow \mathbf{r}_1, \\ 0, & \text{for } \mathbf{r} \rightarrow \mathbf{r}_2, \end{cases} \quad (6.67)$$

Comparing the boundary problem for $\phi(\mathbf{r})$ with that for function $f(\mathbf{r})$, we get $\phi = Vf$. This means that the gradient ∇f , which participates in Eq. (66), is just $(-\mathbf{E}/V) = (-\mathbf{j}/\sigma V)$. Hence the integral in that formula is just $-I/\sigma V = -1/\sigma R_n$, where R_n is the resistance of the Josephson junction in its normal state. As a result, Eq. (66) yields

$$I_c = \frac{\hbar q n_p(T)}{m \sigma} \frac{1}{R_n}, \quad (6.68)$$

showing that the $I_c R_n$ product does not depend on the junction geometry, though it does depend on temperature, vanishing, together with $n_p(T)$, at $T \rightarrow T_c$. (Well below the critical temperature, $I_c R_n$ of a sufficiently short weak links is of the order of $\Delta(0)/e$, i.e. of the order of a few mV.)

Now let us see what happens if a Josephson junction is placed into the gap in a superconductor ring – see Fig. 4c. In this case, we can combine Eqs. (57) and (59), getting

$$I = I_c \sin 2\pi \frac{\Phi}{\Phi_0}. \quad (6.69)$$

This effect of periodic dependence of the current on flux is called the *macroscopic quantum interference*,³³ while the system shown in Fig. 4b, a *superconducting quantum interference device*, abbreviated as *SQUID* (with all letters capital, please :-). The low value of the magnetic flux quantum Φ_0 , and hence the high sensitivity of φ to the magnetic field, allows using SQUIDS as ultrasensitive magnetometers. Indeed, for a superconducting ring of area $\sim 1 \text{ cm}^2$, one period of the change of supercurrent (69) is produced by magnetic field change of the order of 10^{-11} T (10^{-7} Gs), while sensitive electronics allows to measure a tiny fraction of this period – limited by thermal noise at a level of the

³³ The name is due to the deep analogy between this phenomenon and the interference between two waves, to be discussed in detail in Sec. 8.4.

order of a few pT. This sensitivity allows measurements, for example, of the magnetic fields induced by the beating human heart, and even by brain activity, outside of the body.

An important aspect of the quantum interference is the so-called *Aharonov-Bohm (AB) effect*.³⁴ Let the magnetic field lines be limited to the central part of the SQUID ring, so that no appreciable magnetic field ever touches the superconducting ring material. (This may be done experimentally with very good accuracy, for example using high- μ magnetic cores – see their discussion in Sec. 5.6.) As predicted by Eq. (69), and confirmed by several careful experiments carried out in the mid-1960s,³⁵ this restriction does not matter – the interference is observed anyway. This means that not only the magnetic field \mathbf{B} , but also the vector-potential \mathbf{A} represents physical reality, albeit quite a peculiar one – remember the gauge transformation?

Actually, the magnetic flux quantization (55) and the macroscopic quantum interference (69) are not completely different effects, but just two manifestations of the whole group of inter-related macroscopic quantum phenomena. In order to show that, one should note that if critical current I_c (or rather its product by loop's self-inductance L) is high enough, flux Φ in the SQUID loop is due not only to the external magnetic field flux Φ_e , but also has a self-field component - cf. Eq. (5.61):³⁶

$$\Phi = \Phi_{\text{ext}} - LI, \quad \text{where } \Phi_{\text{ext}} \equiv \int_S (\mathbf{B}_{\text{ext}})_n d^2r. \quad (6.70)$$

Now the relation between Φ and Φ_{ext} may be found by solving this equation together with Eq. (69). Figure 5 shows this relation for several values of the dimensionless parameter $\lambda \equiv 2\pi LI_c/\Phi_0$.

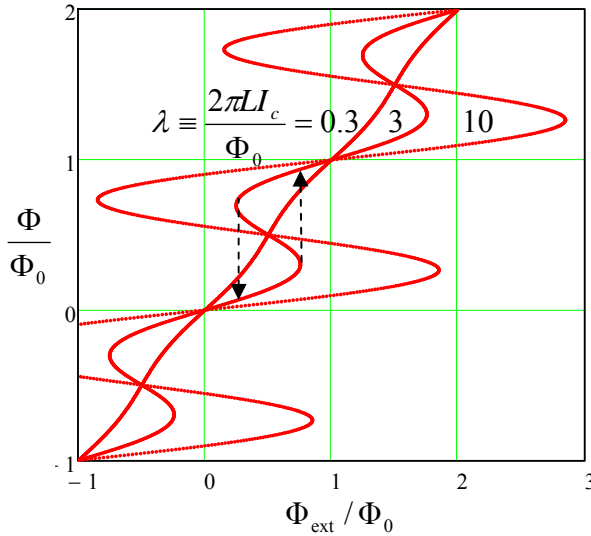


Fig. 6.5. Function $\Phi(\Phi_{\text{ext}})$ for SQUIDs with various values of the normalized LI_c product. Dashed arrows show the flux leaps as the external field is changed. (The branches with $d\Phi/d\Phi_{\text{ext}} < 0$ are unstable.)

³⁴ For a more detailed discussion of the AB effect, which also takes place for single quantum particles, see, e.g., QM Sec. 3.2.

³⁵ Later, similar experiments were carried out with electron beams, and then even with “normal” (meaning non-superconducting) solid-state conducting rings. In this case, the effect is due to interference of the wavefunction of a single charged particle (an electron) with itself, and if of course is much harder to observe than in SQUIDs. In particular, the ring size has to be very small, and temperature low, to avoid “dephasing” effects due to unavoidable interactions of the particles with environment.

³⁶ The sign before LI would be positive, as in Eq. (5.61), if I was the current flowing *into* the inductance. However, in order to keep the sign in Eq. (69) intact, I should mean the current flowing into the Josephson junction, i.e. *from* the inductance, thus changing the sign of the term.

These plots show that if the critical current or (or the inductance) is low, $\lambda \ll 1$, the self-effects are negligible, and the total flux follows the external field (i.e., Φ_{ext}) quite faithfully. However, at $\lambda > 1$, the dependence $\Phi(\Phi_{\text{ext}})$ becomes hysteretic, and at $\lambda \gg 1$ the positive-slope (stable) branches of this function are nearly flat, with the total flux values corresponding to Eq. (55). Thus, a superconducting ring closed by a high- I_c Josephson junction exhibits a nearly-perfect flux quantization.

The self-field effects described by Eq. (70) create certain technical problems for SQUID magnetometry, but they are the basis for one more application of these devices: ultrafast computing. Indeed, Fig. 5 shows that at the values of λ modestly above 1 (e.g., $\lambda \approx 3$), and within a certain range of applied field, the SQUID has two stable flux states that differ by $\Delta\Phi \approx \Phi_0$ and may be used for coding binary 0 and 1. For practical superconductors (like Nb), the time of switching between these states (see dashed arrows in Fig. 4) are of the order of a picosecond, while the energy dissipated at such event may be as low as $\sim 10^{-19}$ J. (This bound is determined not by device's physics, by the fundamental requirement for the energy barrier between the two states to be much higher than the thermal fluctuation energy scale $k_B T$, ensuring a sufficiently long information retention time.) While the picosecond switching speed may be also achieved with some semiconductor devices, the power consumption of the SQUID-based digital devices may be 5 to 6 orders of magnitude lower, enabling VLSI circuits with 100-GHz-scale clock frequencies and manageable power dissipation. Unfortunately, the range of practical application of these *Rapid Single-Flux-Quantum* (RSFQ) logic circuits is still narrow, due to the inconvenience of their deep refrigeration to temperatures below T_c .³⁷

Since we have already got the basic relations (57) and (59) describing the macroscopic quantum phenomena in superconductivity, let me mention in brief two other members of this group, called the *Josephson effects*. Differentiating Eq. (57) over time, and using the Faraday induction law (2), we get³⁸

$$\frac{d\varphi}{dt} = \frac{2e}{\hbar} V. \quad (6.71)$$

Josephson
phase-to-
voltage
relation

This famous *phase-to-voltage relation* should be valid regardless of the way how voltage V has been created,³⁹ so let us apply Eqs. (59) and (71) to the simplest circuit with a non-superconducting source of dc voltage – see Fig. 6.

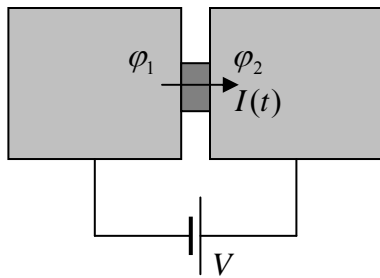


Fig. 6.6. DC-voltage-biased Josephson junction.

³⁷ For more on that technology, see, e.g., the review paper by P. Bunyk *et al.*, *Int. J. High Speed Electron. Syst.* **11**, 257 (2001) and references therein.

³⁸ Since the induced e.m.f. \mathcal{V}_{ind} cannot drop on the superconducting path between the Josephson junction electrodes 1 and 2 (Fig. 3), it should equal to $(-V)$, where V is the voltage across the junction.

³⁹ It may be also obtained from simple Schrödinger equation arguments – see, e.g., QM Sec. 2.2.

If current I is below the critical value,

$$-I_c < I < +I_c, \quad (6.72)$$

Eq. (59) allows phase φ to have a time-independent value

$$\varphi = \arcsin(I/I_c), \quad (6.73)$$

and hence, according to Eq. (71), a vanishing voltage drop across the junction: $V = 0$. This *dc Josephson effect* is not quite surprising – indeed, we have postulated from the very beginning that the Josephson junction may pass a certain supercurrent. Much more fascinating is the so-called *ac Josephson effect* that takes place if voltage across the junction has a nonvanishing average (dc) component V_0 . For simplicity, let us assume that this is the *only* voltage component: $V(t) = V_0 = \text{const}$,⁴⁰ then Eq. (71) may be readily integrated to give $\varphi = \omega_J t + \varphi_0$, where

$$\omega_J \equiv \frac{2e}{\hbar} V_0. \quad (6.74)$$

Josephson
oscillation
frequency

This result, plugged into Eq. (59), shows that supercurrent oscillates,

$$I(\varphi) = I_c \sin(\omega_J t + \varphi_0), \quad (6.75)$$

with the *Josephson frequency* ω_J (74), which is proportional to the applied dc voltage. For practicable voltages, frequency $f_J = \omega_J/2\pi$ corresponds to the GHz or even THz ranges, because the proportionality coefficient in Eq. (74) is very high: $f_J/V_0 = 2e/h \approx 483 \text{ MHz}/\mu\text{V}$.⁴¹ An important experimental fact is the universality of this coefficient. For example, in the mid-1980s, the group led by Prof. J. Lukens of our department proved that this factor is material-independent with the relative accuracy of at least 10^{-15} . Very few experiments, especially in solid state physics, have ever reached such precision.

This fundamental nature of the Josephson voltage-to-frequency relation (74) allows an important application of the ac Josephson effect in metrology. Namely, phase locking the Josephson oscillations with an external microwave signal derived from an atomic frequency standards one can get the most precise dc voltage than from any other source. In NIST and other metrological institutions around the globe, this effect is used for the calibration of simpler “secondary” voltage standards that can operate at room temperature.⁴²

6.5. Inductors, transformers, and ac Kirchhoff laws

Let a *wire coil* (meaning either a single loop illustrated in Fig. 5.4b, or a series of such loops, such as one of the solenoids shown in Fig. 5.6) have size a that satisfies, at frequencies of our interest, the quasistatic limit condition $a \ll \lambda$. Moreover, let the coil’s self-inductance L be much larger than that of the wires connecting it to other components of our system: ac voltage sources, voltmeters, etc. (Since, according to Eq. (5.75), (5.113), L scales as the number N of wire turns squared, this is easier to achieve

⁴⁰ In experiment, this condition is hard to implement, due to relatively high inductance of the current leads providing dc voltage supply. However, these complications do not change the main conclusion of the analysis.

⁴¹ This 1962 prediction by B. Josephson was confirmed experimentally – implicitly (by phase locking of the oscillations with an external oscillator) in 1963, and explicitly (by the detection of microwave radiation) in 1967.

⁴² For more on the Josephson effect and other macroscopic quantum phenomena in superconductivity, see, e.g., Chapters 6 and 7 in the monograph by M. Tinkham, which was cited above.

at $N \gg 1$.) Then in a system consisting of such *lumped induction coils* and external wires (and other circuit elements such as resistors, capacitances, etc.), we may neglect the electromagnetic induction effects everywhere outside the coil, so that the electric field in those external regions is potential. Then the voltage V between coil's terminals may be defined (as in electrostatics) as the difference of values of scalar potential ϕ between the terminals, i.e. as integral

$$V = \int \mathbf{E} \cdot d\mathbf{r} \quad (6.76)$$

between the coil terminals along any path outside the coil. This voltage has to be balanced by the induction e.m.f. (2) in the coil, so that if the Ohmic resistance of the coil is negligible,⁴³ we may write

$$V = \frac{d\Phi}{dt}, \quad (6.77)$$

where Φ is the magnetic flux in the coil. If the flux is due to the current I in the same coil only (i.e. if it is magnetically uncoupled from other coils), we may use Eq. (5.70) to get the well-known relation

$$V = L \frac{dI}{dt}, \quad (6.78)$$

where the compliance with the Lenz sign rule is achieved by selecting the relations between the assumed voltage polarity and current direction as shown in Fig. 7a.

Voltage
drop on
inductance
coil

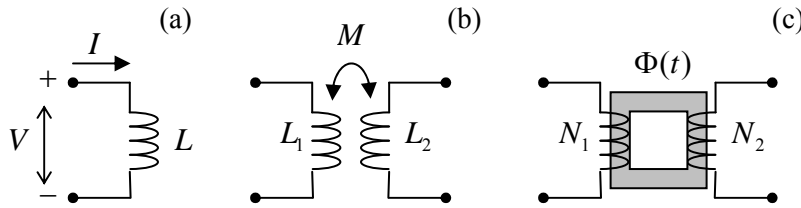


Fig. 6.7. (a) Induction coil, (b) two inductively coupled coils, and (c) an ac transformer.

If similar conditions are satisfied for two magnetically coupled coils (Fig. 6b), then, in Eq. (77), we need to use Eqs. (5.69) instead, getting

$$V_1 = L_1 \frac{dI_1}{dt} + M \frac{dI_2}{dt}, \quad V_2 = L_2 \frac{dI_2}{dt} + M \frac{dI_1}{dt}, \quad (6.79)$$

where the repeating index is dropped for notation simplicity. Such systems of inductively coupled coils have numerous applications in electrical engineering and physical experiment.⁴⁴ Probably the most important is the *ac transformer* (Fig. 6c) where both coils share a common soft-ferromagnetic core. As we already know, such material (with $\mu \gg \mu_0$) tries to not let any magnetic field lines out, so that the magnetic flux $\Phi(t)$ in the core is nearly the same in each of its cross-sections. This gives

$$V_1 \approx N_1 \frac{d\Phi}{dt}, \quad V_2 \approx N_2 \frac{d\Phi}{dt}, \quad (6.80)$$

⁴³ If the resistance is substantial, it may be represented, in calculations, by a separate lumped circuit element (*resistor*) connected in series with the coil.

⁴⁴ Starting from the pioneering experiments by M. Faraday - who invented this device.

where $N_{1,2}$ is the number of wire turns in each coil, so that the voltage ratio is completely determined by N_1/N_2 ratio.

Now we may generalize, to the ac current case, the notion of an *electric circuit*, already discussed in Chapter 4 – see Fig. 4.3 reproduced in Fig. 8a below. Let not only wire inductances but also wire capacitances be negligible in comparison with those of compact (*lumped*) capacitances. Then we may present the circuit as the connection of lumped circuit elements with ideal (voltage- and charge-free wires), with the list of its circuit elements now including not only resistors and current sources (as in the dc case), but also induction coils (including magnetically coupled ones) and capacitors – see Fig. 8b.

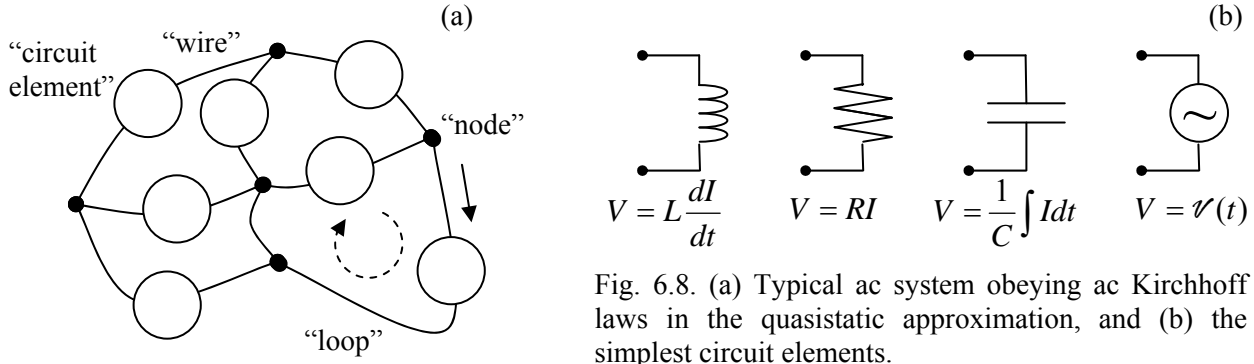


Fig. 6.8. (a) Typical ac system obeying ac Kirchhoff laws in the quasistatic approximation, and (b) the simplest circuit elements.

In the quasistatic limit, current through each wire is conserved, so that the “node rule”, i.e. the 1st Kirchhoff law (4.7),

$$\sum_j I_j = 0. \quad (6.81)$$

remains valid. Also, if the electromagnetic induction effect is restricted to the interiors of lumped induction coils as discussed above, voltage drops V_k across each circuit element may be still presented, just as in dc circuits, as differences of potentials of the adjacent nodes, so that the “loop rule”, i.e. 2nd Kirchhoff law given by Eq. (4.8),

$$\sum_k V_k = 0. \quad (6.82)$$

is also valid.

In contrast to the dc case, Eqs. (81) and (82) are now the (ordinary) differential equations. However, if all circuit elements are linear (as in the examples presented in Fig. 8b), these equations may be readily reduced to linear algebraic equations using the Fourier expansion. (In the most common case of sinusoidal ac sources, the final stage of Fourier series summation is unnecessary.) I do not have time to discuss even the simplest examples of such circuits, such as *LC*, *LR*, *RC*, and *LRC* loops and periodic structures,⁴⁵ but my experience shows that the potential readers of these notes are well familiar with these problems from their undergraduate studies. Let me only emphasize again that the standard ac

⁴⁵ Interestingly, these effects include the wave propagation in periodic *LC* circuits, despite still staying within the quasistatic approximation! However, within this approximation, speed $1/(LC)^{1/2}$ of these waves is much lower than speed $1/(\epsilon\mu)^{1/2}$ of electromagnetic waves in the surrounding medium – see the next chapter.

circuit theory is only valid within the quasistatic limit $a \ll \lambda$, and only under the condition of the electric and magnetic field confinement inside lumped circuit elements.

6.6. Displacement currents

The electromagnetic induction is not the only new effect arising in non-stationary electrodynamics. Indeed, though Eqs. (16) are adequate for the description of quasistatic phenomena, a deeper analysis shows that one of these equations, namely $\nabla \times \mathbf{H} = \mathbf{j}$, cannot be exact. To see that, let us take the divergence of its both sides of this equation:

$$\nabla \cdot (\nabla \times \mathbf{H}) = \nabla \cdot \mathbf{j}. \quad (6.83)$$

But, as the divergence of any curl,⁴⁶ the left hand part should equal zero. Hence we get

$$\nabla \cdot \mathbf{j} = 0. \quad (6.84)$$

This is fine in statics, but in dynamics this equation forbids any charge accumulation, because according to the continuity relation (4.5),

$$\nabla \cdot \mathbf{j} = -\frac{\partial \rho}{\partial t}. \quad (6.85)$$

This discrepancy had been recognized by James Clerk Maxwell who suggested, in 1864, a way out of this contradiction. If we generalize the equation for $\nabla \times \mathbf{H}$ by adding to term \mathbf{j} (that describes real currents) the so-called *displacement current* term,

Displacement
current
density

$$\mathbf{j}_d \equiv \frac{\partial \mathbf{D}}{\partial t}, \quad (6.86)$$

(that of course vanishes in statics), then the equation takes the form

$$\nabla \times \mathbf{H} = \mathbf{j} + \mathbf{j}_d = \mathbf{j} + \frac{\partial \mathbf{D}}{\partial t}. \quad (6.87)$$

In this case, due to equation $\nabla \cdot \mathbf{D} = \rho$, the divergence of the right hand part equals zero due to the continuity equation, and the discrepancy is removed.

This conclusion, and equation (87), are so important that it is worthwhile to have one more look at its derivation using a particular “electrical engineering” model shown in Fig. 8.⁴⁷ Neglecting the fringe field effects, we may use Eq. (4.1) to describe the relation between current I flowing through a wire and the electric charge Q of the capacitor:⁴⁸

⁴⁶ Again, see MA Eq. (11.2) if you need.

⁴⁷ No physicist should be ashamed of doing this. J. C. Maxwell himself has arrived at his equations with a heavy use of *mechanical* engineering arguments. (His main book, *A Treatise of Electricity and Magnetism*, is full of drawings of gears and levers.) More generally, the whole history of science teaches us that snobbishness toward engineering and other “not-a-real-physics” disciplines is a sure way toward producing nothing of either practical value or fundamental importance. In real science, any method leading to novel, correct results should be welcome.

⁴⁸ This is of course just the integral form of the continuity equation (85).

$$\frac{dQ}{dt} = I. \quad (6.88)$$

Now let us consider a closed contour C drawn around the wire. (Solid points in Fig. 9 show the places where the contour intercepts the plane of drawing.) This contour may be seen as either the line limiting surface S_1 (crossed by the wire) or the line limiting surface S_2 (avoiding such crossing by passing through capacitor's gap). Applying the macroscopic Ampère law (5.117) to the former surface, we get

$$\oint_C \mathbf{H} \cdot d\mathbf{r} = \int_{S_1} j_n d^2r = I, \quad (6.89)$$

while for the latter surface the same law gives a different result,

$$\oint_C \mathbf{H} \cdot d\mathbf{r} = \int_{S_2} j_n d^2r = 0, \quad \text{[WRONG!]} \quad (6.90)$$

for the same integral. This is just an integral-form manifestation of the discrepancy outlined above, but it shows clearly how serious the problem is (or rather it was - before Maxwell).

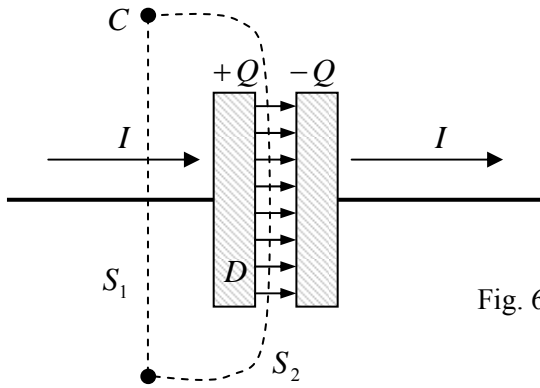


Fig. 6.9. The Ampère law applied to a recharged capacitor.

Now let us see how the introduction of the displacement currents saves the day, considering for the sake of simplicity a plane capacitor of area A , with a constant electrode spacing. In this case, as we already know, the field inside it is uniform, with $D = \sigma$, so that the total capacitor's charge $Q = A\sigma = AD$, and current (88) may be represented as

$$I = \frac{dQ}{dt} = A \frac{dD}{dt}. \quad (6.91)$$

So, instead of Eq. (90), the modified Ampère law gives

$$\oint_C \mathbf{H} \cdot d\mathbf{r} = \int_{S_2} (j_d)_n d^2r = \int_{S_2} \frac{\partial D_n}{\partial t} d^2r = \frac{dD}{dt} A = I, \quad (6.92)$$

i.e. the Ampère integral becomes independent of the choice of the (imaginary!) surface limited by contour C – as it should.

6.7. Finally, the full Maxwell equation system

This is a very special moment in the course: with the displacement current introduction, we have finally arrived at the full set of macroscopic Maxwell equations for time-dependent fields,⁴⁹

Macroscopic
Maxwell
equations

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0, \quad \nabla \times \mathbf{H} - \frac{\partial \mathbf{D}}{\partial t} = \mathbf{j}, \quad (6.93a)$$

$$\nabla \cdot \mathbf{D} = \rho, \quad \nabla \cdot \mathbf{B} = 0, \quad (6.93b)$$

whose validity has been confirmed in by an enormous body of experimental data.⁵⁰ The most striking feature of these equations is that, even in the absence of (local) charges and currents, when all the equations become homogeneous,

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}, \quad \nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t}, \quad (6.94a)$$

$$\nabla \cdot \mathbf{D} = 0, \quad \nabla \cdot \mathbf{B} = 0, \quad (6.94b)$$

they still describe something very non-trivial: electromagnetic waves, including light.⁵¹ Indeed, one can interpret Eqs. (94a) in the following way: the change of magnetic field creates, via the Faraday induction effect, a vortex (divergence-free) electric field, while the dynamics of the electric field, in turn, creates a vortex magnetic field via the Maxwell's displacement currents.

We will carry out a detailed quantitative analysis of the waves in the next chapter, but it is easy (and very instructive) to use the Maxwell equations to estimate their velocity v and the field amplitude ratio E/H in a medium with $\mathbf{D} = \epsilon \mathbf{E}$, $\mathbf{B} = \mu \mathbf{H}$, and $\mathbf{j} = 0$. Indeed, let the solution of these equations, in a uniform, linear medium have a time period T , and hence the wavelength $\lambda = vT$. Then the magnitude of the left-hand part of the first of Eqs. (94a) is of the order of $E/\lambda \sim E/vT$, while that of its right-hand part may be estimated as $B/T = \mu H/T$. Using similar estimates for the second of Eqs. (94a), we arrive at the following two requirements for the E/H ratio:⁵²

$$\frac{E}{H} \sim \mu v \sim \frac{1}{\epsilon v}. \quad (6.95)$$

In order to insure the compatibility of these two relations, the wave speed should satisfy the estimate

$$v \sim \frac{1}{(\epsilon \mu)^{1/2}}, \quad (6.96)$$

reduced to $v \sim 1/(\epsilon_0 \mu_0)^{1/2} \equiv c$ in free space, while the ratio of the electric and magnetic field amplitudes should be of the following order:

⁴⁹ This vector form of the equations, magnificent in its symmetry and simplicity, was developed in 1884-85 by O. Heaviside, with substantial contributions by H. Lorentz. (The original Maxwell's result, circa 1861, looked like a system of 20 equations for Cartesian components of the vector and scalar potentials.)

⁵⁰ Despite numerous efforts, no other corrections (e.g., additional terms) to Maxwell equations have been ever found, and these equations are still considered exact within the range of their validity, i.e. while the electric and magnetic fields may be considered classically. Moreover, even in quantum case, these equations are believed to be *strictly* valid as relations between the Heisenberg operators of the electric and magnetic field.

⁵¹ Let me emphasize that this is only possible due to the “displacement current” term $\partial \mathbf{D}/\partial t$.

⁵² The fact that T cancels shows (or rather hints) that these estimates are valid for waves of arbitrary frequency.

$$\frac{E}{H} \sim \mu v \sim \mu \frac{1}{(\epsilon \mu)^{1/2}} = \left(\frac{\mu}{\epsilon} \right)^{1/2}. \quad (6.97)$$

In the next chapter we will see that these are indeed the *exact* results for a plane electromagnetic wave.

Now let me fulfill the promise given in Sec. 2 and establish the validity limits of the quasistatic approximation (16). For that, let the spatial scale of our system be a , generally unrelated to wavelength $\lambda = vT$, and carry real currents \mathbf{j} producing certain magnetic field H . Then, according to Eqs. (94a), this magnetic field Faraday-induces electric field $E \sim \mu H a / T$, whose displacement currents, in turn, produce an additional magnetic field with magnitude

$$H' \sim \frac{a\epsilon}{T} E \sim \frac{a\epsilon}{T} \frac{\mu a}{T} H \sim \left(\frac{a\lambda}{vT\lambda} \right)^2 H = \left(\frac{a}{\lambda} \right)^2 H. \quad (6.98)$$

Hence, at $a \ll \lambda$, the displacement current effect is indeed negligible.

Before going after the analysis of the full Maxwell equations in particular situations (that will be the main goal of all the next chapters of this course), let us have a look at the energy balance they yield for a certain volume V - that may include both charged particles and the electromagnetic field. Since, according to Eq. (5.10), the magnetic field does no work on charged particles even if they move, the total power \mathcal{P} being transferred from the field to the particles inside the volume is due to the electric field alone:

$$\mathcal{P} = \int_V \mathbf{p} \cdot d^3r, \quad \mathbf{p} = \mathbf{j} \cdot \mathbf{E}, \quad (6.99)$$

where I have used Eq. (4.38). Expressing \mathbf{j} from the corresponding Maxwell equation of system (93), and plugging it into Eq. (99), we get

$$\mathcal{P} = \int_V \left[\mathbf{E} \cdot (\nabla \times \mathbf{H}) - \mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} \right] d^3r. \quad (6.100)$$

Let us pause here for a second, and transform the divergence of vector $\mathbf{E} \times \mathbf{H}$ using the well-known vector algebra identity:⁵³

$$\nabla \cdot (\mathbf{E} \times \mathbf{H}) = \mathbf{H} \cdot (\nabla \times \mathbf{E}) - \mathbf{E} \cdot (\nabla \times \mathbf{H}). \quad (6.101)$$

The last term in the right-hand part of this equation is exactly the first term in the square brackets of Eq. (100), so that we can rewrite that formula as

$$\mathcal{P} = \int_V \left[-\nabla \cdot (\mathbf{E} \times \mathbf{H}) + \mathbf{H} \cdot (\nabla \times \mathbf{E}) - \mathbf{E} \cdot \frac{\partial \mathbf{D}}{\partial t} \right] d^3r. \quad (6.102)$$

However, according to the Maxwell equation for $\nabla \times \mathbf{E}$, it is equal to $-\partial \mathbf{B} / \partial t$, so that the second term in the square brackets of Eq. (102) equals $-\mathbf{H} \cdot \partial \mathbf{B} / \partial t$ and, according to Eq. (5.128), is just the (minus) time derivative of the magnetic energy per unit volume. Similarly, according to Eq. (3.82), the third term under the integral is the minus time derivative of the electric energy per unit volume. Finally, we can use the divergence theorem to transform the integral of the first term to a 2D integral over the surface S

⁵³ See, e.g., MA Eq. (11.7) with $\mathbf{f} = \mathbf{E}$ and $\mathbf{g} = \mathbf{H}$.

limiting volume V . As the result, we get the so-called *Poynting theorem*⁵⁴ for the power balance in the system:

Poynting
theorem

$$\int_V \left(\boldsymbol{\rho} + \frac{\partial u}{\partial t} \right) d^3r + \oint_S S_n d^2r = 0. \quad (6.103)$$

Here u is the density of the total (electric plus magnetic) energy of the electromagnetic field, with

$$\delta u \equiv \mathbf{E} \cdot \delta \mathbf{D} + \mathbf{H} \cdot \delta \mathbf{B}, \quad (6.104a)$$

Electro-
magnetic
energy
density

so that for an isotropic, linear, and dispersion-free medium, with $\mathbf{D}(t) = \epsilon \mathbf{E}(t)$, $\mathbf{B}(t) = \mu \mathbf{H}(t)$,

$$u = \frac{\mathbf{E} \cdot \mathbf{D}}{2} + \frac{\mathbf{H} \cdot \mathbf{B}}{2} = \frac{\epsilon E^2}{2} + \frac{B^2}{2\mu}, \quad (6.104b)$$

and \mathbf{S} is the *Poynting vector* defined as⁵⁵

Poynting
vector

$$\mathbf{S} \equiv \mathbf{E} \times \mathbf{H}. \quad (6.105)$$

The first integral in Eq. (103) is evidently the net change of the energy of the system (particles + field) in unit time, so that the second (surface) integral is certainly the power flowing out from the system through the surface, and it is tempting to interpret the Poynting vector \mathbf{S} locally, as the power flow density at the given point.⁵⁶ In many cases such a local interpretation of vector \mathbf{S} is legitimate; however, in some cases it may lead to wrong conclusions. Indeed, let us consider a simple system shown in Fig. 10: a planar capacitor placed into a static and uniform external magnetic field so that the electric and magnetic fields are mutually perpendicular. In this static situation, no charges are moving, both $\boldsymbol{\rho}$ and $\partial/\partial t$ equal to zero, and there should be no power flow in the system. However, Eq. (105) shows that the Poynting vector is not equal to zero inside the capacitor, being directed as shown in Fig. 10.

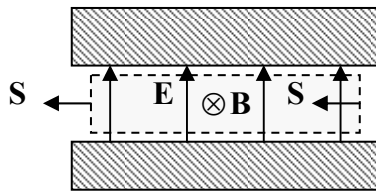


Fig. 6.10. The Poynting vector paradox.

From the point of view of our only unambiguous corollary of the Maxwell equations, Eq. (103), there is no contradiction here, because the fluxes of vector \mathbf{S} through the walls of any volume V , for example the side walls of the volume shown with dashed lines in Fig. 10, are equal and opposite (and they are zero for other faces of this rectilinear volume), so that the total flux of the Poynting vector

⁵⁴ Called after J. Poynting, though this fact was independently discovered by O. Heaviside, while a similar expression for the intensity of mechanical elastic waves had been derived earlier by N. Umov.

⁵⁵ Actually, an addition to \mathbf{S} of the curl of an arbitrary vector function $\mathbf{f}(\mathbf{r}, t)$ does not change Eq. (103). Indeed, we may use the divergence theorem to transform the corresponding change of the surface integral in Eq. (103) to a volume integral of scalar function $\nabla \cdot (\nabla \times \mathbf{f})$ that equals zero at any point – see, e.g., MA Eq. (11.2).

⁵⁶ Later in the course we will show that the Poynting vector is also directly related to the density of momentum of the electromagnetic field.

equals zero, as it should. It is, however, useful to recall this example each time before giving the local interpretation to vector \mathbf{S} .

Finally, to complete the initial discussion of the Maxwell equations,⁵⁷ let us rewrite them in terms of potentials \mathbf{A} and ϕ , because this is more convenient for the solution of some (though not all!) problems. Even when dealing with a more general system (93) of Maxwell equations than before, Eqs. (7) and (5.27),

$$\mathbf{E} = -\nabla\phi - \frac{\partial\mathbf{A}}{\partial t}, \quad \mathbf{B} = \nabla \times \mathbf{A}, \quad (6.106)$$

Electro-
magnetic
field
potentials

are still used as potential definitions. It is straightforward to verify that with these definitions, two homogeneous Maxwell equations (93b) are satisfied automatically. Plugging Eqs. (106) into the inhomogeneous equations (93a), and considering, for simplicity, a linear, uniform medium with frequency-independent ε and μ , we get

$$\begin{aligned} \nabla^2\phi + \frac{\partial}{\partial t}(\nabla \cdot \mathbf{A}) &= -\frac{\rho}{\varepsilon}, \\ \nabla^2\mathbf{A} - \varepsilon\mu \frac{\partial^2\mathbf{A}}{\partial t^2} - \nabla\left(\nabla \cdot \mathbf{A} + \varepsilon\mu \frac{\partial\phi}{\partial t}\right) &= -\mu\mathbf{j}. \end{aligned} \quad (6.107)$$

This is a more complex result than what we would like to get. However, let us select a special gauge that is frequently called (especially for the free space case, when $v = c$) the *Lorenz gauge condition*⁵⁸

$$\nabla \cdot \mathbf{A} + \varepsilon\mu \frac{\partial\phi}{\partial t} = 0, \quad (6.108)$$

Lorenz
gauge
condition

which is a natural generalization of the Coulomb gauge (5.48) for time-dependent phenomena. With this condition, Eqs. (107) are reduced to a simpler, beautifully symmetric form:⁵⁹

$$\begin{aligned} \nabla^2\phi - \frac{1}{v^2} \frac{\partial^2\phi}{\partial t^2} &= -\frac{\rho}{\varepsilon}, \\ \nabla^2\mathbf{A} - \frac{1}{v^2} \frac{\partial^2\mathbf{A}}{\partial t^2} &= -\mu\mathbf{j}. \end{aligned} \quad (6.109)$$

Potential
dynamics

where $v^2 \equiv 1/\varepsilon\mu$.⁶⁰

⁵⁷ We will return to their general discussion (in particular, to the analytical mechanics of the electromagnetic field, and its stress tensor) in Sec. 9.8, after we have got equipped with the special relativity theory.

⁵⁸ This condition, named after *L. Lorenz*, should not be confused with the *Lorentz invariance condition* of the relativity theory, due to *H. Lorentz* (note the names' spelling) – see Sec. 9.4.

⁵⁹ Note that Eqs. (109) are essentially a set of 4 similar equations for 4 scalar functions (namely, ϕ and three Cartesian components of vector \mathbf{A}) and thus clearly invite the 4-component vector formalism of the relativity theory - which will be discussed in Chapter 9.

⁶⁰ Here I have to mention in passing the so-called *Hertz vector potentials* $\mathbf{\Pi}_e$ and $\mathbf{\Pi}_m$ (whose introduction may be traced to at least the 1904 work by E. Whittaker). They may be defined by the following relations:

$$\mathbf{A} = \mu \frac{\partial \mathbf{\Pi}_e}{\partial t} + \mu \nabla \mathbf{\Pi}_m, \quad \phi = -\frac{1}{\varepsilon} \nabla \cdot \mathbf{\Pi}_e,$$

If ϕ and \mathbf{A} depend on just one spatial coordinate, say z , in a region without field sources: $\rho = 0$, $\mathbf{j} = 0$, Eqs. (109) are reduced to the well-known 1D wave equations

$$\begin{aligned}\frac{\partial^2 \phi}{\partial^2 z} - \frac{1}{v^2} \frac{\partial^2 \phi}{\partial t^2} &= 0, \\ \frac{\partial^2 \mathbf{A}}{\partial^2 z} - \frac{1}{v^2} \frac{\partial^2 \mathbf{A}}{\partial t^2} &= 0\end{aligned}\tag{6.110}$$

describing waves propagating with velocity v . Note that due to the definitions of constants ϵ_0 and μ_0 , in free space v is just the speed of light:

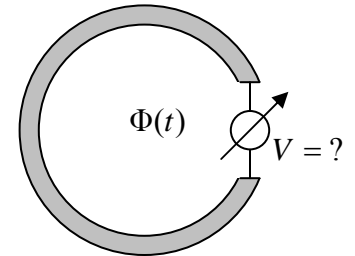
$$v = \frac{1}{(\epsilon_0 \mu_0)^{1/2}} \equiv c.\tag{6.110}$$

Historically, the experimental observation of relatively low-frequency (GHz-scale) electromagnetic waves and the proof that their speed in free space is equal to that of light, was the decisive proof of Maxwell's theory.⁶¹ A detailed study of this most important physical phenomenon is the main goal of the next chapters of this course.

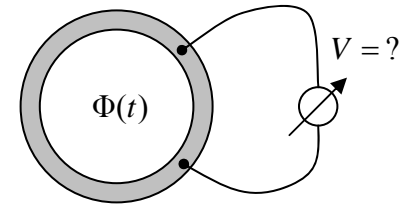
6.8 Exercise problems

6.1. Prove that the electromagnetic induction e.m.f. \mathcal{V}_{ind} in a conducting loop may be measured:

- (i) by measuring the current $I = \mathcal{V}_{\text{ind}}/R$ induced in the closed loop with Ohmic resistance R , or
- (ii) using a voltmeter inserted into the loop – see Fig. on the right.



6.2. The magnetic flux Φ that pierces a plane, round, uniform, resistive ring is being changed in time, while the magnetic field outside of the ring is negligibly low. A voltmeter is connected to a part the ring as shown in Fig. on the right. What would the voltmeter show?



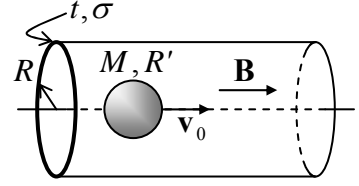
6.3. A weak, uniform magnetic field \mathbf{B} is applied to an axially-symmetric permanent magnet, with a dipole magnetic moment \mathbf{m} directed along its symmetry axis, rapidly rotating about the same axis, with an angular momentum \mathbf{L} . Calculate the electric field resulting from field's application, and formulate the conditions of your result's validity.

which make the Lorenz gauge condition (108) automatically satisfied. These potentials are especially convenient for the solution of problems in which the electromagnetic field is excited by external sources characterized by externally fixed electric and magnetic polarizations \mathbf{P}_{ext} and \mathbf{M}_{ext} - rather than fixed charge and current densities ρ and \mathbf{j} . Indeed, it is straightforward to check that both Π_e and Π_m satisfy equations similar to Eqs. (109), but with the right-hand parts equal to, respectively, $-\mathbf{P}_{\text{ext}}$ and $-\mathbf{M}_{\text{ext}}$.

⁶¹ This was first accomplished in 1886 by H. Hertz, using specially designed electronic circuits and antennas.

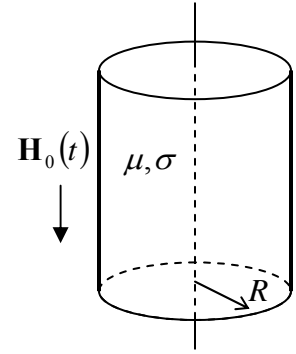
6.4. Use the electromagnetic induction law (5) to derive Eq. (5.128) for the magnetic field energy variation.

6.5. A uniform, static magnetic field \mathbf{B} is applied along the axis of a long round pipe of a radius R , and a very small thickness τ , made of a material with Ohmic conductance σ . A sphere of mass M and radius $R' < R$, made of a linear magnetic with permeability $\mu \gg \mu_0$, is launched, with an initial velocity v_0 , to fly ballistically along pipe's axis – see Fig. on the right. Use the quasistatic approximation to calculate the distance the sphere would pass before it stops. Formulate conditions of validity of your result.



6.6. AC current of frequency ω is being passed through a long uniform wire with a round cross-section of radius R that is comparable with the skin depth δ_s . In the quasistatic approximation, find the current density distribution across the wire. Analyze the limits $R \ll \delta_s$ and $R \gg \delta_s$.

6.7. A very long, round cylinder of radius R , made of a uniform Ohmic conductor with conductivity σ and magnetic permeability μ , has been placed into a uniform ac magnetic field $\mathbf{H}_{\text{ext}} = \mathbf{H}_0 \cos \omega t$, directed along its symmetry axis – see Fig. on the right. Calculate the spatial distribution of the magnetic field's amplitude, and in particular its value on cylinder's axis. Spell out the last result in the limits of relatively small and large R .



Hint: As shortcuts, you are welcome to reuse parts of the solution of the previous problem.

6.8.* Define and calculate an appropriate spatial-temporal Green's function for Eq. (20), and use this function to analyze the dynamics of propagation of the external magnetic field, suddenly turned on at $t = 0$ and then left constant:

$$H(x=0, t) = \begin{cases} 0, & \text{at } t < 0, \\ H_0, & \text{at } t > 0, \end{cases}$$

into an Ohmic conductor occupying half-space $x > 0$ – see Fig. 6.2.

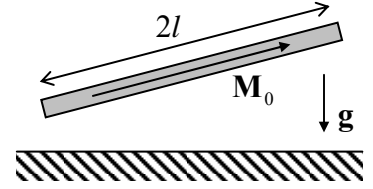
Hint: Try to use a function proportional to $\exp\{-(x - x')^2/2(\delta x)^2\}$, with a suitable time dependence of parameter δx , and a properly selected pre-exponential factor.

6.9. Solve the previous problem using the variable separation method, and compare the results.

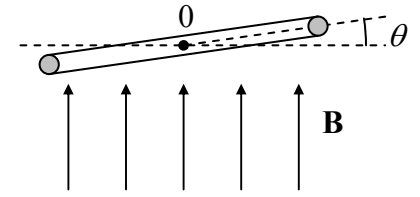
6.10. A small, planar wire loop, carrying current I , is located far from a plane surface of a superconductor. Within the “macroscopic” description of superconductivity ($\mathbf{B} = 0$), find:

- the energy of the loop-superconductor interaction,
- the force and torque acting on the loop,
- the distribution of supercurrents on the superconductor surface.

6.11. A straight, uniform magnet of length $2l$, cross-section area $A \ll l^2$, and mass m , with a permanent longitudinal magnetization M_0 , is placed over a horizontal surface of a superconductor – see Fig. on the right. Within the macroscopic model of the Meissner-Ochsenfeld effect, find the stable equilibrium position of the magnet.



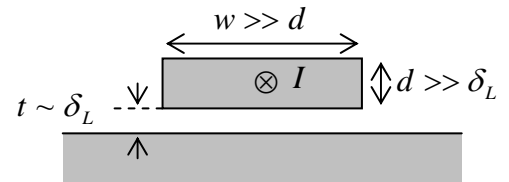
6.12. A plane superconducting wire loop, of area A and inductance L , may rotate, without friction, about a horizontal axis O (in Fig. on the right, perpendicular to the plane of drawing) passing through its center of mass. Initially the loop was horizontal (with $\theta = 0$), and carried supercurrent I_0 in such direction that its magnetic dipole vector was directed down. Then a uniform magnetic field \mathbf{B} , directed vertically up, was applied. Find all possible equilibrium positions (angles θ) of the loop, analyze their stability, and give a physical interpretation of the results.



6.13. Use the London equation to analyze the penetration of external magnetic field into a thin ($t \sim \delta_L$), planar superconductor film whose plane is parallel to the field.

6.14. Use the London equation to calculate the distribution of supercurrent density \mathbf{j} across the circular cross-section (with radius $R \sim \delta_L$) of a long, straight superconducting wire carrying dc current I .

6.15.* Use the London equation to calculate the inductance (per unit length) of a long, uniform superconducting strip placed close to the surface of a similar superconductor – see Fig. on the right, which shows the structure's cross-section.

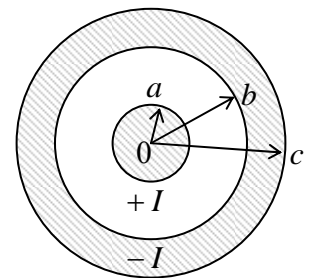


Hint: Start from thinking how is the supercurrent distributed along the surfaces of the strip and the bulk superconductor.

6.16. Analyze the magnetic field shielding by a superconducting film of small thickness $t \ll \delta_L$, by calculating the penetration of the field induced by current I flowing in a thin wire which runs parallel to a wide, plane thin film, at distance $d \gg t$ from it, into the half-space behind the film.

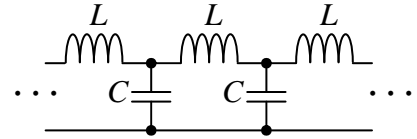
6.17. Calculate the self-inductance of a superconducting cable with a round cross-section (see Fig. on the right) in the following limits:

- (i) $\delta_L \ll a, b, c - b$, and
- (ii) $a \ll \delta_L \ll b, c - b$.



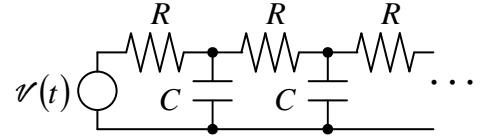
6.18. Use Eqs. (59) and (71) to calculate the energy of a Josephson junction, and the full energy of the SQUID shown in Fig. 4c.

6.19. Analyze the possibility of wave propagation in a long, uniform chain of lumped inductances and capacitances – see Fig. on the right.



Hint: Readers without prior experience with electromagnetic wave analysis may like to use a substantial analogy between this effect and mechanical waves in a 1D chain of elastically coupled particles.⁶²

6.20. A sinusoidal e.m.f. of amplitude V_0 and frequency ω is applied to an end of a long chain of similar, lumped resistors and capacitors (see Fig. on the right). Calculate the law of decay of the rf oscillation amplitude in the chain.

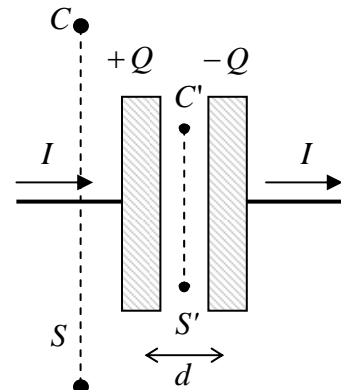


6.21. Calculate the pressure exerted by the magnetic field \mathbf{B} inside a magnetic-free solenoid of length l , cross-section area $A \ll l^2$ and N turns, on its “walls” (windings), and the force exerted by the field on solenoid’s ends. Give a physical interpretation of the direction of these forces.

6.22. In Sec. 6 we have seen that the displacement current concept allows one to generalize the Ampère law to time-dependent processes as

$$\oint_C \mathbf{H} \cdot d\mathbf{r} = I_S + \frac{\partial}{\partial t} \int_S D_n d^2r.$$

We also have seen that such generalization makes $\oint \mathbf{H} \cdot d\mathbf{r}$ over the contour C , which was shown in Fig. 9 (see also Fig. on the right), independent of the choice of surface S limited by the contour. However, it may look like the situation is different for contour C' drawn inside the capacitor. Indeed, if contour’s radius ρ is much larger than the capacitor’s thickness d , the magnetic field \mathbf{H} , created by the linear current I of the contour line is virtually the same as that of a continuous wire, and hence integral $\oint \mathbf{H} \cdot d\mathbf{r}$ along contour C' is apparently the same as that along contour C , while the magnetic flux $\int D_n d^2r$ through the surface S' limited by contour C' is evidently smaller, while $I_{S'} = I_S = 0$, so that the above equation seems invalid. Resolve the paradox, for simplicity considering an axially-symmetric system.



⁶² See, e.g., CM Sec. 5.3.

Chapter 7. Electromagnetic Wave Propagation

This (long!) chapter focuses on the most important effect that follows from the time-dependent Maxwell equations, namely the electromagnetic waves, at this stage avoiding a discussion of their origin, i.e. radiation. I start from the simplest, plane waves in a uniform and isotropic media. The next step is a discussion non-uniform systems, in particular those with sharp boundaries between different materials, which bring in such new effects as wave reflection and refraction. Then I will proceed to the structure of electromagnetic waves propagating along various long, cylindrical structures, called transmission lines - such as coaxial cables, waveguides, and optical fibers. In the end of the chapter, electromagnetic oscillations in final-length fragments of such lines, serving as resonators, are also discussed.

7.1. Plane waves

Let us start from considering a spatial region that does not contain field sources ($\rho = 0$, $\mathbf{j} = 0$), and is filled with a linear, uniform, isotropic medium, which obeys Eqs. (3.38) and (5.110):

$$\mathbf{D} = \varepsilon \mathbf{E}, \quad \mathbf{B} = \mu \mathbf{H}. \quad (7.1)$$

Moreover, let us assume for a minute that these material equations hold for all frequencies of interest. As was already shown in Sec. 6.7, in this case the Lorenz gauge condition (6.108) allows the Maxwell equations to be recast into wave equations (6.110) for the vector and scalar potentials. However, for most our purposes it is more convenient to use directly the homogeneous Maxwell equations (6.94) for the electric and magnetic fields - which are independent of the gauge choice. After the elementary elimination of \mathbf{D} and \mathbf{B} using Eq. (1),¹ these equations take a simple, symmetric form

$$\nabla \times \mathbf{E} + \mu \frac{\partial \mathbf{H}}{\partial t} = 0, \quad \nabla \times \mathbf{H} - \varepsilon \frac{\partial \mathbf{E}}{\partial t} = 0, \quad (7.2a)$$

$$\nabla \cdot \mathbf{E} = 0, \quad \nabla \cdot \mathbf{H} = 0. \quad (7.2b)$$

Now, taking the curl ($\nabla \times$) of each of Eqs. (2a), and using the vector algebra identity (5.31), whose first term, for both \mathbf{E} and \mathbf{H} , vanishes due to Eqs. (2b), we get similar wave equations for the electric and magnetic fields:

$$\left(\nabla^2 - \frac{1}{v^2} \frac{\partial^2}{\partial t^2} \right) \mathbf{E} = 0, \quad \left(\nabla^2 - \frac{1}{v^2} \frac{\partial^2}{\partial t^2} \right) \mathbf{H} = 0, \quad (7.3)$$

where parameter v is defined by relation

$$v^2 \equiv \frac{1}{\varepsilon \mu}. \quad (7.4)$$

with $v^2 = 1/\varepsilon_0 \mu_0 \equiv c^2$ in free space.

¹ Though \mathbf{B} rather than \mathbf{H} is the actual (microscopically-averaged) magnetic field, it is mathematically more convenient (just as in Sec. 6.2) to use the latter vector in our current discussion, because at sharp media boundaries, \mathbf{H} obeys the boundary condition (5.118) similar to that for \mathbf{E} – see Eq. (3.47).

Two vector equations (3) are of course six similar equations for three Cartesian components of two vectors \mathbf{E} and \mathbf{H} . Each of these equations allows, in particular, the following solution,

$$f = f(z - vt), \quad (7.5) \quad \text{Plane wave}$$

where z is the Cartesian coordinate along a certain (arbitrary) direction \mathbf{n} . This solution describes a specific type of a *wave*, i.e. a certain field pattern moving, without deformation, along axis z , with velocity v . According to Eq. (5), each variable f has the same value in each plane perpendicular to the direction \mathbf{n} of wave propagation, hence the name – *plane wave*.

According to Eqs. (2), the independence of the wave *equations* (3) for vectors \mathbf{E} and \mathbf{H} does not mean that their plane-wave *solutions* are independent. Indeed, plugging solution (5) into Eqs. (2a), we get

$$\mathbf{H} = \frac{\mathbf{n} \times \mathbf{E}}{Z}, \quad \text{i.e. } \mathbf{E} = Z \mathbf{H} \times \mathbf{n}, \quad (7.6) \quad \text{Relation between the fields}$$

where constant Z is defined as

$$Z \equiv \frac{E}{H} = \left(\frac{\mu}{\epsilon} \right)^{1/2}. \quad (7.7) \quad \text{Wave impedance}$$

The vector relation (6) means, first of all, that vectors \mathbf{E} and \mathbf{H} are perpendicular not only to vector \mathbf{n} (such waves are called *transverse*), but also to each other (Fig. 1) - at any point of space and at any time instant.

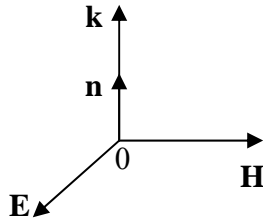


Fig. 7.1. Field vectors in a plane electromagnetic wave propagating along direction \mathbf{n} .

Second, the field magnitudes are related by constant Z , called the *wave impedance* of the medium. Very soon we will see that the wave impedance plays a pivotal role in many problems, in particular at the wave reflection from the interface between two media. Since the dimensionality of E , in SI units, is V/m, and that of H is A/m, Eq. (7) shows that Z has the dimensionality of V/A, i.e. ohms (Ω).² In particular, in free space,

$$Z = Z_0 \equiv \left(\frac{\mu_0}{\epsilon_0} \right)^{1/2} = 4\pi \times 10^{-7} c \approx 377 \Omega. \quad (7.8) \quad \text{Wave impedance of free space}$$

Now plugging Eq. (6) into Eqs. (6.104b) and (6.105), we get:

$$u = \epsilon E^2 = \mu H^2, \quad (7.9a) \quad \text{Wave's energy per unit volume}$$

² In Gaussian units, E and H have the same dimensionality (in particular, in a free-space wave, $E = H$), making the (very useful) notion of the wave impedance less manifestly exposed - and in some textbooks not mentioned at all.

Wave's
power
per unit
area

$$\mathbf{S} \equiv \mathbf{E} \times \mathbf{H} = \mathbf{n} \frac{E^2}{Z} = \mathbf{n} Z H^2, \quad (7.9b)$$

so that, according to Eqs. (4) and (7), wave's energy and power densities are universally related as

$$\mathbf{S} = \mathbf{n} u v. \quad (7.9c)$$

In the view of the Poynting vector paradox discussed in Sec. 6.7 (see Fig. 6.10), one may wonder whether this expression may be interpreted as the actual density of power flow. In contrast to the static situation shown in Fig. 6.7, that limits the electric and magnetic fields to a vicinity of their sources, waves may travel far from them. As a result, they can form *wave packets* of finite length in free space – see Fig. 2.

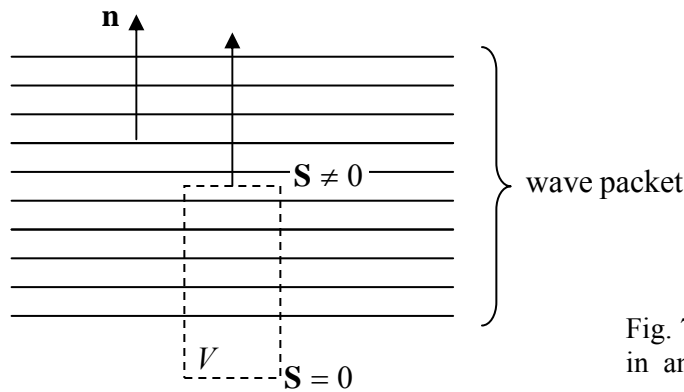


Fig. 7.2. Interpreting the Poynting vector in an electromagnetic wave.

Let us apply the Poynting theorem (6.103) to the cylinder shown by dashed lines in Fig. 2, with one lid inside the wave packet, and another lid in the region already passed by the wave. Then, according to Eq. (6.103), the rate of change of the full energy \mathcal{E} inside the volume is $d\mathcal{E}/dt = -SA$ (where A is the lid area), so that S may be indeed interpreted as the power flow (per unit area) from the volume. Making a reasonable assumption that the finite length of a sufficiently long wave packet does not affect the physics inside it, we may indeed interpret the \mathbf{S} given by Eq. (9) as the power flow density inside a plane electromagnetic wave.

As we will see later in this chapter, the free-space value Z_0 of the wave impedance, given by Eq. (8), establishes the scale of wave impedances of virtually all wave transmission lines, so we may use it and Eq. (9) to get some sense of how different are the electric and magnetic field amplitudes in the waves, on the scale of typical electrostatics and magnetostatics experiments. For example, according to Eqs. (9), a wave of a modest intensity $S = 1 \text{ W/m}^2$ (the power density we get from a usual electric bulb a few meters away from it) has $E \sim (SZ_0)^{1/2} \sim 20 \text{ V/m}$, quite comparable with the dc field created by an AA battery right outside it. On the other hand, the wave's magnetic field $H = (S/Z_0)^{1/2} \approx 0.05 \text{ A/m}$. For this particular case, the relation following from Eqs. (1), (4), and (7),

$$B = \mu H = \mu \frac{E}{Z} = \mu \frac{E}{(\mu/\epsilon)^{1/2}} = (\epsilon\mu)^{1/2} E = \frac{E}{v}, \quad (7.10)$$

gives $B = \mu_0 H = E/c \sim 7 \times 10^{-8} \text{ T}$, i.e. a magnetic field thousand times less than the Earth field, and about 8 orders of magnitude lower than the field of a typical permanent magnet. A possible interpretation of this huge difference is that the scale of magnetic fields $B \sim E/c$ in the waves is “normal” for

electromagnetism, while that of permanent magnet fields is abnormally high, because they are due to the ferromagnetic alignment of electron spins, essentially quantum objects – see the discussion in Sec. 5.5.

As soon as ε and μ are simple constants, wave speed v is also constant, and Eq. (5) is valid for an arbitrary function f - defined by either initial or boundary conditions. In plain English, a medium with frequency-independent ε and μ supports propagation of plane waves with an arbitrary waveform without either decay (*attenuation*) or deformation (*dispersion*). However, for any real medium but pure vacuum, this approximation is valid only within limited frequency intervals. We will discuss the effects of attenuation and dispersion in the next section and see that all our prior results remain valid even in that general case, provided that we limit them to single-frequency (i.e. sinusoidal, or *monochromatic*) waves. Such waves may be most conveniently presented as³

$$f = \text{Re} \left[f_{\omega} e^{i(kz - \omega t)} \right], \quad (7.11)$$

Mono-
chromatic
wave

where f_{ω} is the *complex amplitude* of the wave, and k is its *wave number* (the magnitude of *wave vector* $\mathbf{k} \equiv \mathbf{n}k$), sometimes also called the *spatial frequency*. The last term is justified by the fact, evident from Eq. (11), that k is related to the wavelength λ exactly as the usual (“temporal”) frequency ω is related to the time period T :

$$k = \frac{2\pi}{\lambda}, \quad \omega = \frac{2\pi}{T}. \quad (7.12)$$

Spatial and
temporal
frequencies

Requiring Eq. (11) to be a particular form of Eq. (5), i.e. the argument $(kz - \omega t) \equiv k[z - (\omega/k)t]$ to be proportional to $(z - vt)$, so that $\omega/k = v$, we see that the wave number should equal

$$k = \frac{\omega}{v} = (\varepsilon\mu)^{1/2} \omega, \quad (7.13)$$

Dispersion
relation

showing that in this “dispersion-free” case the *dispersion relation* $\omega(k)$ is linear.

Now note that Eq. (6) does not claim mean vectors \mathbf{E} and \mathbf{H} retain their direction in space. (The simple case when they do is called the *linear polarization* of the wave.) Indeed, nothing in the Maxwell equations prevents, for example, joint rotation of this pair of vectors around the fixed vector \mathbf{n} , while still keeping all these three vectors perpendicular to each other at all times. An arbitrary rotation law, or even an arbitrary constant frequency of such rotation, however, would violate the single-frequency (monochromatic) character of the elementary sinusoidal wave (11). In order to understand what is the most general type of polarization the wave may have without violating that condition, let us present two Cartesian components of one of these vectors (say, \mathbf{E}) along any two fixed axes x and y , perpendicular to each other and axis z (i.e. vector \mathbf{n}), in the same form as used in Eq. (11):

$$E_x = \text{Re} \left[E_{\omega x} e^{i(kz - \omega t)} \right], \quad E_y = \text{Re} \left[E_{\omega y} e^{i(kz - \omega t)} \right]. \quad (7.14)$$

In order to keep the wave monochromatic, complex amplitudes $E_{\omega x}$ and $E_{\omega y}$ must be constant; however, they may have different magnitudes and an arbitrary phase shift between them.

³ Due to the linearity of Eqs (2), operator Re in Eq. (11) may be ignored until the end of almost any calculation. Because of that, the exponential presentation of monochromatic variables is more convenient than manipulation with sine and cosine functions. (See also CM Sec. 4.1.)

In the simplest case when the arguments of the complex amplitudes are equal,

$$E_{\omega x,y} = |E_{\omega x,y}| e^{i\varphi}. \quad (7.15)$$

the real field components have the same phase:

$$E_{x,y} = |E_{\omega x,y}| \cos(kz - \omega t + \varphi), \quad (7.16)$$

so that their ratio is constant in time – see Fig. 3a. This means that the wave is linearly polarized, within the plane defined by relation

$$\tan \theta = \frac{|E_{\omega y}|}{|E_{\omega x}|}. \quad (7.17)$$

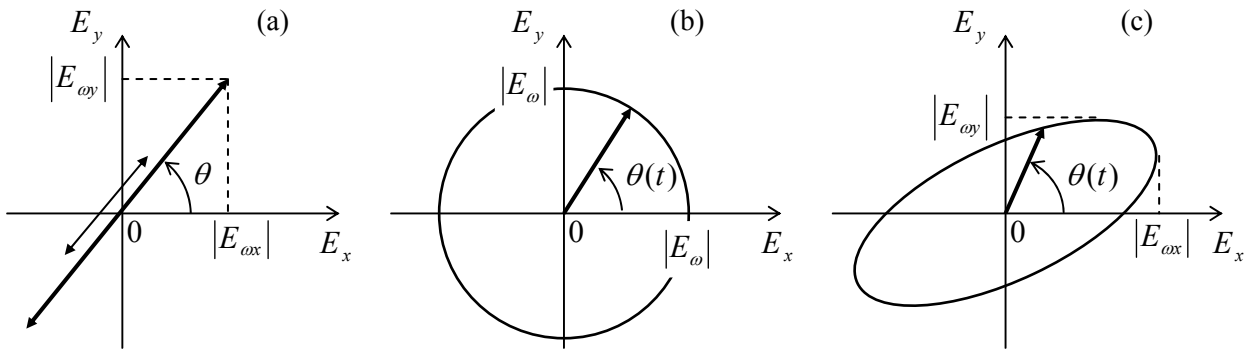


Fig. 7.3. Time evolution of the electric field vector in (a) linearly-polarized, (b) circularly-polarized, and (c) elliptically-polarized waves.

Another simple case is when the moduli of the complex amplitudes $E_{\omega x}$ and $E_{\omega y}$ are equal, but their phases are shifted by $+\pi/2$ or $-\pi/2$:

$$E_{\omega x} = |E_\omega| e^{i\varphi}, \quad E_{\omega y} = |E_\omega| e^{i(\varphi \pm \pi/2)}. \quad (7.18)$$

In this case

$$E_x = |E_\omega| \cos(kz - \omega t + \varphi), \quad E_y = |E_\omega| \cos\left(kz - \omega t + \varphi \pm \frac{\pi}{2}\right) = \mp |E_\omega| \sin(kz - \omega t + \varphi). \quad (7.19)$$

This means that on the $[x, y]$ plane, the end of vector \mathbf{E} moves, with wave's frequency ω , either clockwise or counterclockwise around a circle – see Fig. 3b:

$$\theta(t) = \mp(\omega t - \varphi). \quad (7.20)$$

Such waves are called *circularly-polarized*.⁴ These particular solutions of the Maxwell equations are very convenient for quantum electrodynamics, because single electromagnetic field quanta with a

⁴ In the convention that dominates research and engineering literature (but unfortunately is not universal), the wave is called *right-polarized* (RP) if it is described by the lower sign in Eqs. (18)-(20), and *left-polarized* (LP) in the opposite case. Another popular term for these cases is the “waves of negative / positive *helicity*”.

certain (positive or negative) spin direction may be considered as elementary excitations of the corresponding circularly-polarized wave. (This fact does not exclude, from the quantization scheme, waves of other polarizations, because any monochromatic wave may be presented as a linear combination of two circularly-polarized waves with opposite helicities, just as Eqs. (14) present it as a linear combination of two linearly-polarized waves.)

Finally, in the general case of arbitrary complex amplitudes $E_{\omega x}$ and $E_{\omega y}$, the electric field vector end moves along an ellipse on the $[x, y]$ plane (Fig. 3c), such wave is called *elliptically polarized*. The eccentricity and orientation of the ellipse are completely described by one complex number, the ratio $E_{\omega x}/E_{\omega y}$, i.e. two real numbers: $|E_{\omega x}/E_{\omega y}|$ and $\varphi = \arg(E_{\omega x}/E_{\omega y})$.

The same information may be expressed via four so-called *Stokes parameters* s_0, s_1, s_2, s_3 , which are popular in optics because they may be used for the description of not only completely coherent waves that are discussed here, but also of partly coherent or even fully incoherent waves - including the *natural light* emitted by thermal sources like our Sun. In contrast to the notion of coherent waves whose complex amplitudes are considered deterministic numbers, the instant amplitudes of incoherent waves should be treated as stochastic variables.⁵

7.2. Attenuation and dispersion

Now let me show that *any* linear, isotropic medium may be characterized, by complex, frequency-dependent electric permittivity $\varepsilon(\omega)$ and magnetic permeability $\mu(\omega)$. Indeed, starting from electric effects, the electric polarization of realistic elementary dipoles of the medium cannot follow the applied electric field instantly, if the field frequency ω is comparable with those of the internal processes - say, transitions between atomic energy levels. Let us consider the most general law of time evolution of polarization $P(t)$ for the case of arbitrary applied electric field $E(t)$,⁶ but for a sufficiently dilute medium, so that the local electric field \mathbf{E}_{ef} (3.63), acting on each elementary dipole, is essentially the microscopically-averaged field \mathbf{E} .⁷ Then, due to the linear superposition principle, $P(t)$ should be a linear sum (integral) of the values of $E(t')$ at all previous moments of time, $t' < t$, weighed by some function of t and t' :

$$P(t) = \int_{-\infty}^t E(t') G(t, t') dt'. \quad (7.21)$$

Temporal
Green's
function

The condition $t' < t$, which is implied by this relation, expresses a key principle of physics, the *causal relation* between a cause (in our case, the electric field applied to each dipole) and its effect (the

⁵ For further reading about the Stokes parameters, as well as about many optics topics I will not have time to cover (especially the geometrical optics and the diffraction-imposed limits on optical imaging resolution), I can recommend the classical text by M. Born *et al.*, *Principles of Optics*, 7th ed., Cambridge U. Press, 1999.

⁶ In an isotropic media, vectors \mathbf{E} , \mathbf{P} , and hence $\mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}$, are all parallel, and for the notation simplicity I will drop the vector sign. I am also assuming that \mathbf{P} at any point \mathbf{r} is only dependent on the electric field at the same point, and hence drop term ikz from the exponent's argument. This assumption is valid if wavelength λ is much larger than the elementary media dipole size a . In most systems of interest, the scale of a is atomic ($\sim 10^{-10}$ m), so that the last approximation is valid up to very high frequencies, $\omega \sim c/a \sim 10^{18} \text{ s}^{-1}$, corresponding to hard X-rays.

⁷ Note that this condition (which excludes, in particular, the molecular-field effects discussed in Sec. 3.5) is not mentioned in most E&M textbooks. If the molecular fields are important, Eq. (21) and its corollaries are only valid for the relation between \mathbf{P} and the effective local electric field \mathbf{E}_{ef} .

polarization it creates). Function $G(t, t')$ is called the *temporal Green's function* for the electric polarization.⁸ In order to understand its physical sense, let us consider the case when the applied field $E(t)$ is a very short pulse at $t = t_0$, that may be approximated with the Dirac's delta-function:

$$E(t) = \delta(t - t_0). \quad (7.22)$$

Then Eq. (21) yields just $P(t) = G(t, t_0)$, showing that the Green's function is just the polarization at moment t , created by a unit δ -functional pulse of the applied field at moment t' (Fig. 4). Thus, the temporal G is the exact time analog of the spatial Green's functions $G(\mathbf{r}, \mathbf{r}')$ we have already studied in the electrostatics – see Sec. 2.7.

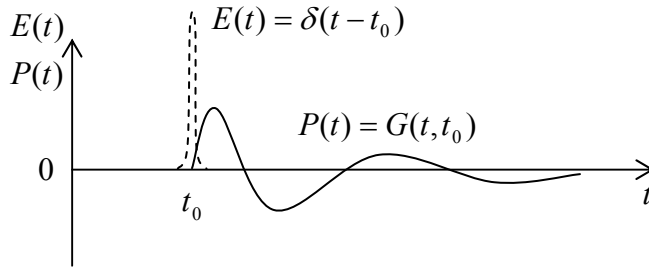


Fig. 7.4. Temporal Green's function for electric polarization (schematically).

What are the general properties of the temporal Green's function? First, the function is evidently real, since the dipole moment \mathbf{p} and hence polarization $\mathbf{P} = n\mathbf{p}$ are real by the definition – see Eq. (3.6). Next, for systems without infinite internal memory, G should tend to zero at $t - t' \rightarrow \infty$, although the type of this approach (e.g., whether function G oscillates approaching zero) depends on the medium. Finally, if parameters of the medium do not change in time, the polarization response to an electric field pulse should depend not on its absolute timing, but only on the time difference $\theta \equiv t - t'$ between the pulse and observation instants:

$$P(t) = \int_{-\infty}^t E(t') G(t - t') dt' = \int_0^{\infty} E(t - \theta) G(\theta) d\theta. \quad (7.23)$$

For a sinusoidal waveform, $E(t) = \text{Re} [E_{\omega} e^{-i\omega t}]$, this equation yields

$$P(t) = \text{Re} \int_0^{\infty} E_{\omega} e^{-i\omega(t-\theta)} G(\theta) d\theta = \text{Re} \left\{ \left[E_{\omega} \int_0^{\infty} G(\theta) e^{i\omega\theta} d\theta \right] e^{-i\omega t} \right\}. \quad (7.24)$$

The expression in square brackets is of course nothing more than the complex amplitude P_{ω} of the polarization. This means that though even if the static relation (3.35) $P = \chi_e \epsilon_0 E$ is invalid for an arbitrary time-dependent process, we may still keep its Fourier analog,

$$P_{\omega} = \chi_e(\omega) \epsilon_0 E_{\omega}, \quad \text{with } \chi_e(\omega) \equiv \frac{1}{\epsilon_0} \int_0^{\infty} G(\theta) e^{i\omega\theta} d\theta, \quad (7.25)$$

⁸ A discussion of the temporal Green's functions in application to classical oscillations may be also found in CM Sec. 4.1.

for each sinusoidal component of the process, using it as the definition of the frequency-dependent electric susceptibility $\chi_e(\omega)$. Similarly, the frequency-dependent electric permittivity may be defined using the Fourier analog of Eq. (3.38):

$$D_\omega \equiv \varepsilon(\omega)E_\omega. \quad (7.26)$$

Then, according to Eq. (3.36), the permittivity is related to the temporal Green's function by the usual Fourier transform:

$$\varepsilon(\omega) \equiv \varepsilon_0 + \frac{P_\omega}{E_\omega} = \varepsilon_0 + \int_0^\infty G(\theta)e^{i\omega\theta} d\theta. \quad (7.27)$$

Complex
electric
permittivity

It is evident from this expression that $\varepsilon(\omega)$ may be complex,

$$\varepsilon(\omega) = \varepsilon'(\omega) + i\varepsilon''(\omega), \quad \varepsilon'(\omega) = \varepsilon_0 + \int_0^\infty G(\theta)\cos\omega\theta d\theta, \quad \varepsilon''(\omega) = \int_0^\infty G(\theta)\sin\omega\theta d\theta, \quad (7.28)$$

and that its real part $\varepsilon'(\omega)$ is always an even function of frequency, while the imaginary part $\varepsilon''(\omega)$ is an odd function of ω .

Absolutely similar arguments show that the linear magnetic properties may be characterized with complex, frequency-dependent permeability $\mu(\omega)$. Now rewriting Eqs. (1) for the complex amplitudes of the fields at a particular frequency, we may repeat all calculations of Sec. 1, and verify that all its results are valid for monochromatic waves even for a dispersive (but necessarily linear!) medium. In particular, Eqs. (7) and (13) now become

$$Z(\omega) = \left(\frac{\mu(\omega)}{\varepsilon(\omega)} \right)^{1/2}, \quad k(\omega) = \omega[\varepsilon(\omega)\mu(\omega)]^{1/2}, \quad (7.28)$$

so that the wave impedance and wave number may be both complex functions of frequency.

This fact has important consequences for the electromagnetic wave propagation. First, plugging the presentation of the complex wave number as the sum of its real and imaginary parts, $k(\omega) \equiv k'(\omega) + ik''(\omega)$, into Eq. (11):

$$f = \operatorname{Re} \left\{ f_\omega e^{i[k(\omega)z - \omega t]} \right\} = e^{-k''(\omega)z} \operatorname{Re} \left\{ f_\omega e^{i[k'(\omega)z - \omega t]} \right\}, \quad (7.29)$$

we see that $k''(\omega)$ describes the rate of wave *attenuation* in the medium at frequency ω .⁹ Second, if the waveform is not sinusoidal (and hence should be presented as a sum of several/many sinusoidal components), the frequency dependence of $k'(\omega)$ provides for wave *dispersion*, i.e. the waveform deformation at the propagation, because the propagation velocity (4) of component waves is now different.¹⁰

⁹ It may be tempting to attribute this effect to wave *absorption*, i.e. the dissipation of the wave's energy, but we will see very soon that wave attenuation may be also due to effects different from absorption.

¹⁰ The reader is probably familiar with the most noticeable effect of the dispersion, namely the difference between that *group velocity* $v_{\text{gr}} \equiv d\omega/dk'$, giving the speed of the envelope of a wave packet with a narrow frequency spectrum, and the *phase velocity* $v_{\text{ph}} \equiv \omega/k'$ of the component waves. The second-order dispersion effect, proportional to $d^2\omega/dk'^2$, leads to the deformation (gradual broadening) of the envelope itself. Following tradition,

Let us consider a simple but very important *Lorentz oscillator model* of a dispersive medium.¹¹ In dilute atomic or molecular systems (including gases), electrons respond to the external electric field especially strongly when frequency ω is close to certain eigenfrequencies ω_j corresponding to the spectrum of quantum transitions of a single atom/molecule. An approximate, phenomenological description of this behavior may be obtained from a classical model of several externally-driven harmonic oscillators with finite damping. For an oscillator, driven by electric field's force $F(t) = qE(t)$, we can write the 2nd Newton law as

$$m(\ddot{x} + 2\delta\dot{x} + \omega_0^2 x) = qE(t), \quad (7.30)$$

where ω_0 is the own frequency of the oscillator, and δ its damping coefficient. For a sinusoidal field, $E(t) = \text{Re} [E_\omega \exp\{-i\omega t\}]$, we can look for a particular, *forced-oscillation* solution in a similar form $x(t) = \text{Re} [x_\omega \exp\{-i\omega t\}]$.¹² Plugging this solution into Eq. (30), we can readily find the complex amplitude of these oscillations:

$$x_\omega = \frac{q}{m} \frac{E_\omega}{(\omega_0^2 - \omega^2) - 2i\omega\delta}. \quad (7.31)$$

Using this result to calculate the complex amplitude of the dipole moment as $p_\omega = qx_\omega$, and then the electric polarization $P_\omega = np_\omega$ of a dilute medium with n independent oscillators for unit volume, for its frequency-dependent permittivity (27) we get

$$\varepsilon(\omega) = \varepsilon_0 + \frac{nq^2}{m} \frac{1}{(\omega_0^2 - \omega^2) - 2i\omega\delta}. \quad (7.32)$$

This result may be readily generalized to the case when the system has several types of oscillators with different eigenfrequencies:

$$\varepsilon(\omega) = \varepsilon_0 + n \frac{q^2}{m} \sum_j \frac{f_j}{(\omega_j^2 - \omega^2) - 2i\omega\delta_j}, \quad (7.33)$$

where $f_j \equiv n_j/n$ is the fraction of oscillators with eigenfrequency ω_j , so that the sum of all f_j equals 1. Figure 5 shows a typical behavior of the real and imaginary parts of the complex dielectric constant, described by Eq. (33), as functions of frequency. The effect of oscillator resonances is clearly visible, and dominates the media response at $\omega \approx \omega_j$, especially in the case of low damping, $\delta_j \ll \omega_j$. Note that in the low-damping limit, the imaginary part of the dielectric constant ε'' , and hence the wave attenuation k'' , are negligibly small at all frequencies besides small vicinities of frequencies ω_j , where derivative $d\varepsilon'(\omega)/d\omega$ is negative.¹³ Thus, for a system of for weakly-damped oscillators, Eq. (33) may be approximated, at most frequencies, as a sum of odd singularities (“poles”):

these effects are discussed in more detail in the quantum-mechanics part of my lecture notes (QM Sec. 2.1), because they are the crucial factor of Schrödinger's wave mechanics. (See also CM Sec. 5.3.)

¹¹ This example is focused on the frequency dependence of ε , because electromagnetic waves interact with “usual” media via their electric field much more than via the magnetic field. However, as will be discussed in Sec. 7, forgetting about the possible dispersion of $\mu(\omega)$ might result in missing some remarkable opportunities for manipulating the waves.

¹² If this point is not absolutely clear, please see CM Sec. 3.1.

¹³ In optics, such behavior is called the *anomalous dispersion*.

$$\varepsilon(\omega) \approx \varepsilon_0 + n \frac{q^2}{2m} \sum_j \frac{f_j}{\omega_j - \omega}, \quad \text{for } \delta_j \ll |\omega - \omega_j| \ll |\omega_j - \omega_{j'}|. \quad (7.34)$$

This result is especially important because, according to quantum mechanics,¹⁴ Eq. (34) is also valid for a set of non-interacting, similar quantum systems (whose dynamics may be completely different from that of a harmonic oscillator!), provided that ω_j are replaced with frequencies of possible quantum interstate transitions, and coefficients f_j are replaced with the so-called *oscillator strengths* of the transitions - which obey the same *sum rule*, $\sum_j f_j = 1$.

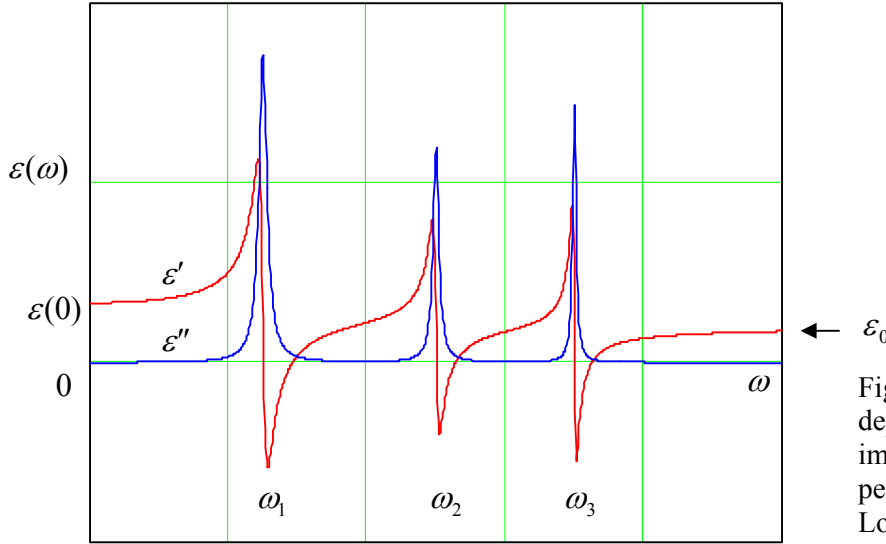


Fig. 7.5. Typical frequency dependence of the real and imaginary parts of the electric permittivity of a media in the Lorentz oscillator model.

At $\omega \rightarrow 0$, the imaginary part of the permittivity also vanishes (for any δ_j), while its real part approaches its electrostatic (“dc”) value

$$\varepsilon(0) = \varepsilon_0 + q^2 \sum_j \frac{n_j}{m_j \omega_j^2}. \quad (7.35)$$

Note that according to Eq. (30), the denominator in Eq. (35) is just the effective spring constant $\kappa_j = m_j \omega_j^2$ of the j^{th} oscillator, so that the oscillator masses m_j as such are actually (and quite naturally) not involved in the static dielectric response.

In the opposite limit $\omega \gg \omega_j$, δ_j , permittivity (33) also becomes real, and may be presented as

$$\varepsilon(\omega) = \varepsilon_0 \left(1 - \frac{\omega_p^2}{\omega^2} \right), \quad \text{where } \omega_p^2 \equiv \frac{q^2}{\varepsilon_0} \sum_j \frac{n_j}{m_j}. \quad (7.36) \quad \varepsilon(\omega) \text{ in plasma}$$

The last result is very important, because it is also valid at *all* frequencies if all ω_j and δ_j vanish, i.e. for a gas of free charged particles, in particular for *plasmas* – ionized atomic gases, with negligible collision effects. (This is why the parameter ω_p defined by Eq. (36) is called the *plasma frequency*.) Typically, the plasma as a whole is neutral, i.e. the density n of positive atomic ions is equal to that of

¹⁴ See, e.g., QM Chapters 5 and 9.

the free electrons. Since the ratio n_j/m_j for electrons is much higher than that for ions, the general formula (36) for the plasma frequency is usually well approximated by the following simple expression:

$$\omega_p^2 \equiv \frac{ne^2}{\varepsilon_0 m_e}. \quad (7.37)$$

This expression has a simple physical sense: the effective spring constant $\kappa_{\text{ef}} = m_e \omega_p^2 = ne^2/\varepsilon_0$ describes the Coulomb force that appears when the electron subsystem of a plasma is shifted, as a whole, from its positive-ion subsystem, thus violating the electroneutrality. Indeed, consider such a small shift, Δx , perpendicular to the plane surface of a broad, plane slab filled with plasma. The uncompensated charges, with equal and opposite surface densities $\sigma = \mp en\Delta x$, that appear at the slab surfaces, create inside the it, according to Eq. (2.3), a uniform electric field $E_x = en\Delta x/\varepsilon_0$. This field exerts force $eE = (ne^2/\varepsilon_0) \Delta x$ on each positively charged ion. According to the 3rd Newton law, the ions pull each electron back to its equilibrium position with the equal and opposite force $F = -eE = - (ne^2/\varepsilon_0) \Delta x$, justifying the above expression for κ_{ef} . Hence it is not surprising that $\varepsilon(\omega)$ described by the first of Eqs. (36) turns into zero at $\omega = \omega_p$: at this resonance frequency, finite free oscillations of charge (and hence of $D = \varepsilon E$) do not require a finite force (and hence E).

The behavior of electromagnetic waves in a medium that obeys Eq. (36), is very remarkable. If the wave frequency ω is above ω_p , the dielectric constant and hence the wave number (28) are positive and real, and waves propagate without attenuation, following the dispersion relation,

$$k(\omega) = \omega [\varepsilon(\omega) \mu_0]^{1/2} = \frac{1}{c} (\omega^2 - \omega_p^2)^{1/2}, \quad (7.38)$$

which is shown in Fig. 6. (As we will see later in this chapter, many wave transmission systems obey such dispersion law as well.)

Plasma
dispersion
relation

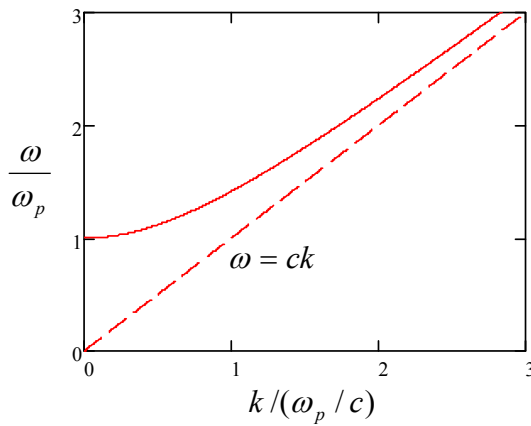


Fig. 7.6. Plasma dispersion law (solid line) in comparison with the linear dispersion of the free space (dashed line).

At $\omega \rightarrow \omega_p$ the wave number k tends to zero. Beyond that point (at $\omega < \omega_p$), we still can use Eq. (38), but it is more instrumental to rewrite it in the mathematically equivalent form

$$k(\omega) = \frac{i}{c} (\omega_p^2 - \omega^2)^{1/2} = \frac{i}{\delta}, \quad \text{where } \delta \equiv \frac{c}{(\omega_p^2 - \omega^2)^{1/2}}. \quad (7.39)$$

According to Eq. (29), this means that the electromagnetic field exponentially decreases with distance:

$$f = \text{Re } f_\omega e^{i(kz - \omega t)} = \exp\left\{-\frac{z}{\delta}\right\} \text{Re } f_\omega e^{-i\omega t}. \quad (7.40)$$

Does this mean that the wave is being absorbed in the plasma? Answering this question is a good pretext to calculate the time average of the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$ of a monochromatic electromagnetic wave in an *arbitrary* dispersive (but still linear!) medium. First, let us spell out fields' time dependence:

$$E(t) = \text{Re}[E_\omega(z)e^{-i\omega t}] = \frac{1}{2}[E_\omega e^{-i\omega t} + \text{c.c.}], \quad H(t) = \text{Re}[H_\omega(z)e^{-i\omega t}] = \frac{1}{2}\left[\frac{E_\omega}{Z(\omega)}e^{-i\omega t} + \text{c.c.}\right]. \quad (7.41)$$

Now, a straightforward calculation yields¹⁵

$$\bar{S} = \overline{E(t)H(t)} = \frac{E_\omega E_\omega^*}{4} \left[\frac{1}{Z(\omega)} + \frac{1}{Z^*(\omega)} \right] = \frac{E_\omega E_\omega^*}{2} \text{Re} \frac{1}{Z(\omega)} \equiv \frac{|E_\omega|^2}{2} \text{Re} \left(\frac{\varepsilon(\omega)}{\mu(\omega)} \right)^{1/2}. \quad (7.42)$$

Let us apply this important general formula to our simple model of plasma at $\omega < \omega_p$. In this case $\mu(\omega) = \mu_0$, i.e. is positive and real, while $\varepsilon(\omega)$ is real and negative, so that $1/Z(\omega) = [\varepsilon(\omega)/\mu(\omega)]^{1/2}$ is purely imaginary, and the average Poynting vector (42) vanishes. This means that energy, on the average, does not flow along axis z – as it would if it was absorbed in plasma. As we will see in the next section, waves with $\omega < \omega_p$ are rather *reflected* from plasma's boundary, without energy loss. Note that in the limit $\omega \ll \omega_p$, Eq. (39) yields

$$\delta \rightarrow \frac{c}{\omega_p} = \left(\frac{c^2 \varepsilon_0 m_e}{ne^2} \right)^{1/2} = \left(\frac{m_e}{\mu_0 ne^2} \right)^{1/2}. \quad (7.43)$$

But this is just a particular case (for $q = e$ and $\mu = \mu_0$) of the expression (6.38) that we have derived for the depth of magnetic field penetration into a lossless (collision-free) conductor in the quasistatic approximation. We see again that, as was already discussed in Sec. 6.7, that approximation (in which we neglect the displacement currents) gives an adequate description of the time-dependent phenomena at $\omega \ll \omega_p$, i.e. at $\delta \ll c/\omega = 1/k = \lambda/2\pi$.

There are two most important examples of plasmas. For the Earth's ionosphere, i.e. the upper part of the atmosphere that is almost completely ionized by the UV and X-ray components of Sun's radiation, the maximum value of n , reached at about 300 km over the Earth surface, is between 10^{10} and 10^{12} m^{-3} (depending on the time of the day and Sun's activity), so that that the maximum plasma frequency (37) is between 1 and 10 MHz. This is much higher than the particle's reciprocal collision time τ^{-1} , so that Eq. (36) gives a very good description of plasma's electric polarization. The effect of reflection of waves with $\omega < \omega_p$ from the ionosphere enables long-range (over-the-globe) radio communications and broadcasting at the so-called *short waves*, with frequencies of the order of 10 MHz.

¹⁵ For an arbitrary plane wave the total average power flow may be calculated as an integral of Eq. (42) over all frequencies. By the way, combining this integral and the Poynting theorem (6.103), one can also prove the following interesting expression for the average electromagnetic energy density in an arbitrary dispersive (but linear and isotropic) medium:

$$\bar{u} = \frac{1}{2} \int_{\omega} \left[\frac{d(\omega\varepsilon)}{d\omega} E_\omega E_\omega^* + \frac{d(\omega\mu)}{d\omega} H_\omega H_\omega^* \right] d\omega.$$

Such waves may propagate in the flat channel formed by the Earth surface and the ionosphere, reflected repeatedly by these “walls”. Unfortunately, due to the random variations of Sun’s activity, and hence ω_p , such natural communication channel is not too reliable, and in our age of fiber optics cables its practical importance is diminishing.

Another important example of plasmas is free electrons in metals and other conductors. For a typical metal, n is of the order of $10^{23} \text{ cm}^{-3} = 10^{29} \text{ m}^{-3}$, so that Eq. (37) yields $\omega_p \sim 10^{16} \text{ s}^{-1}$. Note that this value of ω_p is somewhat higher than mid-optical frequencies ($\omega \sim 3 \times 10^{15} \text{ s}^{-1}$). This explains why planar, even, clean metallic surfaces, such as aluminum and silver films used in mirrors, are so shiny: at these frequencies the permittivity is almost exactly real and negative, leading to light reflection, with very little absorption. However, the considered model, which neglects electron scattering, becomes inadequate at lower frequencies, $\omega\tau \sim 1$.

A phenomenological way of extending the model by account of scattering is to take, in Eq. (33), the lowest eigenfrequency ω_j to be equal zero (to describe free electrons), while keeping the damping coefficient δ_0 of this mode finite, to account for their energy loss due to scattering. Then Eq. (33) is reduced to

$$\varepsilon_{\text{ef}}(\omega) = \varepsilon_{\text{opt}}(\omega) + \frac{n_0 q^2}{m} \frac{1}{-\omega^2 - 2i\omega\delta_0} = \varepsilon_{\text{opt}}(\omega) + \frac{i}{\omega} \frac{n_0 q^2}{2\delta_0 m} \frac{1}{1 - i\omega/2\delta_0}, \quad (7.44)$$

where response $\varepsilon_{\text{opt}}(\omega)$ at high (in practice, optical) frequencies is still given by Eq. (33), but now with $j \neq 0$.

Result (44) allows for a simple interpretation. To show that, let us incorporate into our calculations the Ohmic conduction, generalizing Eq. (4.7) as $\mathbf{j}_\omega = \sigma(\omega)\mathbf{E}_\omega$ to account for the possible frequency dependence of the Ohmic conductivity. Plugging this relation into the Fourier image of the relevant Maxwell equation, $\nabla \times \mathbf{H}_\omega = \mathbf{j}_\omega - i\omega\mathbf{D}_\omega = \mathbf{j}_\omega - i\omega\varepsilon(\omega)\mathbf{E}_\omega$, we get

$$\nabla \times \mathbf{H}_\omega = [\sigma(\omega) - i\omega\varepsilon(\omega)]\mathbf{E}_\omega. \quad (7.45)$$

This relation shows that for a sinusoidal process, the addition of the Ohmic current density \mathbf{j}_ω to the displacement current density is equivalent to addition of $\sigma(\omega)$ to $-i\omega\varepsilon(\omega)$, i.e. to the following change of the ac electric permittivity:¹⁶

$$\varepsilon(\omega) \rightarrow \varepsilon_{\text{ef}}(\omega) \equiv \varepsilon_{\text{opt}}(\omega) + i \frac{\sigma(\omega)}{\omega}. \quad (7.46)$$

Now the comparison of Eqs. (44) and (46) shows that they coincide if we take

$$\sigma(\omega) = \frac{n_0 q^2 \tau}{m_0} \frac{1}{1 - i\omega\tau} = \sigma(0) \frac{1}{1 - i\omega\tau}, \quad (7.47)$$

where the dc conductivity $\sigma(0)$ is described by the Drude formula (4.13), and the phenomenologically introduced coefficient δ_0 is associated with $1/2\tau$. Relation (47), which is frequently called the

¹⁶ Alternatively, according to Eq. (45), it is possible (and in infrared spectroscopy, conventional) to attribute the ac response of a medium at *all* frequencies to effective complex conductivity $\sigma_{\text{ef}}(\omega) = \sigma(\omega) - i\omega\varepsilon(\omega) = -i\omega\varepsilon_{\text{ef}}(\omega)$.

generalized (or “ac”, or “rf”) *Drude formula*,¹⁷ gives a very reasonable (semi-quantitative) description of the ac conductivity of many metals almost all the way up to optical frequencies.

7.3. Kramers-Kronig relations

The results for the simple model of dispersion, discussed in the last section, imply that the frequency dependences of the real (ε') and imaginary (ε'') parts of the permittivity are not quite independent. For example, let us have one more look at the resonance peaks in Fig. 5. Each time the real part drops with frequency, $d\varepsilon'/d\omega < 0$, its imaginary part ε'' has a positive peak. R. de L. Kronig in 1926 and H. A. Kramers in 1927 independently showed that this is not an occasional coincidence pertinent only to the Lorentz oscillator model. Moreover, the full knowledge of function $\varepsilon'(\omega)$ allows one to *calculate* function $\varepsilon''(\omega)$, and vice versa. The reason is that both these functions are always related to a single real function $G(\theta)$ by Eqs. (28).

To derive the Kramers-Kronig relations, let us consider Eq. (27) on the complex frequency plane, $\omega \rightarrow \omega' + i\omega''$:

$$f(\omega) \equiv \varepsilon(\omega) - \varepsilon_0 = \int_0^\infty G(\theta) e^{i\omega\theta} d\theta = \int_0^\infty G(\theta) e^{i\omega'\theta} e^{-\omega''\theta} d\theta. \quad (7.48)$$

For all stable physical systems, $G(\theta)$ has to be finite for all important values of the integration variable ($\theta > 0$), and tend to zero at $\theta \rightarrow 0$ and $\theta \rightarrow \infty$. Because of that, and thanks to factor $e^{-\omega''\theta}$, the expression under the integral tends to zero at $|\omega| \rightarrow \infty$ in all upper half-plane ($\omega'' \geq 0$). As a result, we may claim that the complex-variable function $f(\omega)$ is analytical in that half-plane, and allows us to apply to it the *Cauchy integral* formula¹⁸

$$f(\omega) = \frac{1}{2\pi i} \oint_C f(\Omega) \frac{d\Omega}{\Omega - \omega}, \quad (7.49)$$

with the integration contour of the form shown in Fig. 7, with radius R of the larger semicircle tending to infinity, and radius r that of the smaller semicircle (about the singular point $\Omega = \omega$) tending to zero.

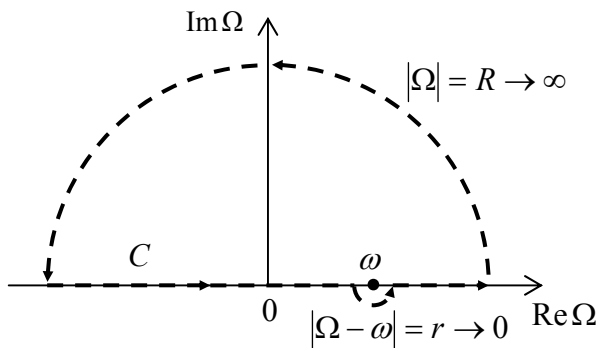


Fig. 7.7. Integration path C used in the Cauchy integral formula to derive the Kramers-Kronig dispersion relations.

¹⁷ It may be also derived from the Boltzmann kinetic equation in the so-called relaxation-time approximation (RTA) – see, e.g., SM Sec. 6.2.

¹⁸ See, e.g., MA Eq. (15.2).

Due to the exponential decay of $|f(\Omega)|$ at $|\Omega| \rightarrow \infty$, the contribution to the integral from the larger semicircle vanishes,¹⁹ while the contribution from the small semicircle, where $\Omega = \omega + r \exp\{i\varphi\}$, with $-\pi \leq \varphi \leq 0$, is

$$\lim_{r \rightarrow 0} \frac{1}{2\pi i} \int_{\Omega=\omega+r \exp\{i\varphi\}} f(\Omega) \frac{d\Omega}{\Omega - \omega} = \frac{f(\omega)}{2\pi i} \int_{-\pi}^0 \frac{ir \exp\{i\varphi\} d\varphi}{r \exp\{i\varphi\}} = \frac{f(\omega)}{2\pi} \int_{-\pi}^0 d\varphi = \frac{1}{2} f(\omega). \quad (7.50)$$

As a result, for our contour C , Eq. (49) yields

$$f(\omega) = \lim_{r \rightarrow 0} \frac{1}{2\pi i} \left(\int_{-\infty}^{\omega-r} + \int_{\omega+r}^{+\infty} \right) f(\Omega) \frac{d\Omega}{\Omega - \omega} + \frac{1}{2} f(\omega). \quad (7.51)$$

Such an integral, excluding a symmetric infinitesimal vicinity of the pole singularity, is called the *principal value* of the (formally, diverging) integral from $-\infty$ to $+\infty$, and is denoted by letter P before it.²⁰ Using this notation, subtracting $f(\omega)/2$ from both parts of Eq. (48), and multiplying them by 2, we get

$$f(\omega) = \frac{1}{\pi i} P \int_{-\infty}^{+\infty} f(\Omega) \frac{d\Omega}{\Omega - \omega}. \quad (7.52)$$

Now plugging into this complex equality the polarization-related difference $f(\omega) \equiv \varepsilon(\omega) - \varepsilon_0$ in the form $[\varepsilon'(\omega) - \varepsilon_0] + i[\varepsilon''(\omega)]$, and requiring both real and imaginary components of both parts of Eq. (52) to be equal separately, we get the famous *Kramers-Kronig dispersion relations*

$$\varepsilon'(\omega) = \varepsilon_0 + \frac{1}{\pi} P \int_{-\infty}^{+\infty} \varepsilon''(\Omega) \frac{d\Omega}{\Omega - \omega}, \quad \varepsilon''(\omega) = -\frac{1}{\pi} P \int_{-\infty}^{+\infty} [\varepsilon'(\Omega) - \varepsilon_0] \frac{d\Omega}{\Omega - \omega}. \quad (7.53)$$

Kramers-
Kronig
dispersion
relations

Now we may use the already mentioned fact that $\varepsilon'(\omega)$ is always an even, while $\varepsilon''(\omega)$ an odd function of frequency, to rewrite these relations in the following form

$$\varepsilon'(\omega) = \varepsilon_0 + \frac{2}{\pi} P \int_0^{+\infty} \varepsilon''(\Omega) \frac{\Omega d\Omega}{\Omega^2 - \omega^2}, \quad \varepsilon''(\omega) = -\frac{2\omega}{\pi} P \int_0^{+\infty} [\varepsilon'(\Omega) - \varepsilon_0] \frac{d\Omega}{\Omega^2 - \omega^2}, \quad (7.54)$$

which is more convenient for most applications, because it involves only physical (positive) frequencies.

Though the Kramers-Kronig relations are “global” in frequency, in certain cases they allow an approximate calculation of dispersion from experimental data for absorption, collected even in a limited frequency range. For example, if a medium has a sharp absorption peak at some frequency ω_j , we may approximate it as

$$\varepsilon''(\omega) \approx c \delta(\omega - \omega_j) + \text{a more smooth function of } \omega, \quad (7.55)$$

and the first of Eqs. (54) immediately gives

¹⁹Strictly speaking, this also requires $|f(\Omega)|$ to decrease faster than Ω^{-1} at the real axis (at $\Omega'' = 0$), but due to nonvanishing inertia of charged particles, this requirement is fulfilled for all realistic models of dispersion – see, e.g., Eq. (36).

²⁰ I am typesetting this symbol in a Roman font, to exclude any possibility of its confusion with media’s polarization.

$$\varepsilon'(\omega) \approx \varepsilon_0 + \frac{2c}{\pi} \frac{\omega_j}{\omega_j^2 - \omega^2} + \text{another smooth function of } \omega, \quad (7.56)$$

Dispersion
near an
absorption
line

thus predicting the anomalous dispersion near such a point. This calculation shows that such behavior observed in the Lorentz oscillator model (Fig. 5) is by no means occasional or model-specific.

Let me emphasize again that the general, and hence very powerful Kramers-Kronig relations hinge on the causal, linear relation (21) between polarization $P(t)$ with the electric field $E(t')$, but not on much else. This is why such relations are also valid for similar causal relations in other fields of physics.²¹

7.4. Reflection

The most important new effect arising in nonuniform media is wave *reflection*. Let us start its discussion from the simplest case of a plane electromagnetic wave that is normally incident on an interface between two uniform, linear, isotropic media.

If the interface is an ideal mirror, the description of reflection is very simple. Indeed, let us assume that one of the two media (say, located at $z > 0$, see Fig. 8) cannot sustain any electric field at all:

$$E|_{z \geq 0} = 0. \quad (7.57)$$

This condition is evidently incompatible with the single traveling wave (5). However, this solution may be readily corrected using the fact that the dispersion-free 1D wave equation,

$$\left(\frac{\partial^2}{\partial z^2} - \frac{1}{v^2} \frac{\partial^2}{\partial t^2} \right) E = 0, \quad (7.58)$$

supports waves, propagating, with the same speed, in opposite directions. As a result, the following linear superposition of two such waves,

$$E|_{z \leq 0} = f(z - vt) - f(-z - vt), \quad (7.59)$$

²¹ In this context, it is important to remember that a simply-looking relation between Fourier amplitudes of certain variables, such as $\mathbf{D}_\omega = \varepsilon(\omega)\mathbf{E}_\omega$, still does not imply the causal relationship between them. This means that the Kramers-Kronig relations are not necessarily valid for either functions $\varepsilon(\omega)$ and $\mu(\omega)$, or their reciprocals, of an arbitrary medium. Indeed, since any Green's function describing a causal relationship has to tend to zero at small times $\theta \equiv t - t'$ (because no system may respond to an external force instantly), its Fourier image has to tend to zero at $\omega \rightarrow \pm \infty$. This is certainly true, for example, for function $f(\omega) \equiv \varepsilon(\omega) - \varepsilon_0$ given by Eq. (32) describing a dilute electric medium, but not for its inverse $1/f(\omega) \propto (\omega^2 - \omega_0^2) - 2i\delta\omega$, which diverges at large frequencies. As another example, since in a dilute linear medium the magnetic response should be due to a causal relation between the average magnetic field \mathbf{B} (cause) and magnetization \mathbf{M} (effect), whose Fourier images are related as $\mathbf{M}_\omega = \chi_m(\omega)\mathbf{H}_\omega = [1/\mu_0 - 1/\mu(\omega)]\mathbf{B}_\omega$, the Kramers-Kronig relations may be expected to be valid for function $f'(\omega) \equiv 1/\mu_0 - 1/\mu(\omega)$, but not for $\mu(\omega)$ or even $[\mu(\omega) - \mu_0]$. Unfortunately, magnetic susceptibility dispersion studies were started just recently, mostly in the context of the negative refractivity effects – see Sec. 5 below, and I am not aware of any convincing discussion of this issue even in research literature (leave alone textbooks :-).

satisfies both the equation and the boundary condition (57), for an arbitrary function f . The second term in Eq. (59) may be interpreted as the *total reflection* of the *incident wave* described by its first term, in this case with the change of electric field's sign. By the way, since vector \mathbf{n} of the reflected wave is opposite to that incident one (see arrows in Fig. 1), Eq. (6) shows that the magnetic field of the wave does not change its sign at the reflection:

$$H|_{z \leq 0} = \frac{1}{Z} [f(z - vt) + f(-z - vt)]. \quad (7.60)$$

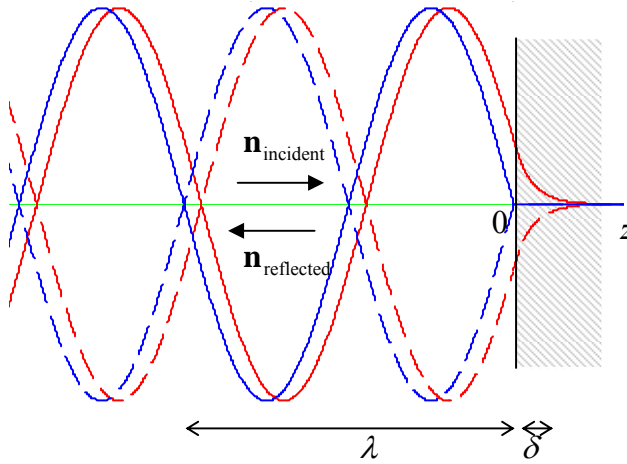


Fig. 7.8. Spatial dependence of electric field at the reflection of a sinusoidal wave from a perfect conductor: the real pattern (red lines) and the crude, ideal-mirror approximation (blue lines). Dashed lines show the patterns after a half-period time delay ($\omega\Delta t = \pi$).

Blue lines in Fig. 8 show the resulting pattern (59) for the simplest, sinusoidal waveform

$$E|_{z \leq 0} = \text{Re} [E_\omega e^{i(kz - \omega t)} - E_\omega e^{i(-kz - \omega t)}]. \quad (7.61a)$$

Depending on convenience in a particular context, this pattern may be legitimately interpreted either as a superposition (61a) of two traveling waves or a single *standing wave*,

$$E|_{z \leq 0} = -2 \text{Im} (E_\omega e^{-i\omega t}) \sin kz = 2 \text{Re} (i E_\omega e^{-i\omega t}) \sin kz, \quad (7.61b)$$

in which the electric and magnetic field oscillate with the phase shifts by $\pi/2$ both in time and space:

$$H|_{z \leq 0} = \text{Re} \left[\frac{E_\omega}{Z} e^{i(kz - \omega t)} + \frac{E_\omega}{Z} e^{i(-kz - \omega t)} \right] = 2 \text{Re} \left(\frac{E_\omega}{Z} e^{-i\omega t} \right) \cos kz. \quad (7.62)$$

As the result of this shift, the time average of the Poynting vector's magnitude,

$$S(z, t) = EH = \frac{1}{Z} \text{Re} [E_\omega^2 e^{-2i\omega t}] \sin 2kz, \quad (7.63)$$

equals zero, showing that at the total reflection there is no *average* power flow. (This is natural, because the perfect mirror can neither transmit the wave nor absorb it.) However, Eq. (63) shows that the standing wave provides local oscillations of energy, transferring it periodically between the concentrations of the electric and magnetic fields, separated by distance $\Delta z = \pi/2k = \lambda/4$.

For the case of the sinusoidal waves, the reflection effects may be readily explored even for the more general case of dispersive and/or lossy media in which $\epsilon(\omega)$ and $\mu(\omega)$, and hence the wave vector

Wave's
total
reflection

$k(\omega)$ and wave impedance $Z(\omega)$, defined by Eqs. (28), are certain complex functions of frequency. The “only” new factors we have to account for is that in this case the reflection may not be full, and that inside the second media we have to use the traveling-wave solution as well. Both these factors may be taken care of by looking for the solution of our boundary problem in the form

$$E|_{z \leq 0} = \text{Re} \left[E_\omega \left(e^{ik_- z} + R e^{-ik_- z} \right) e^{-i\omega t} \right], \quad E|_{z \geq 0} = \text{Re} \left[E_\omega T e^{ik_+ z} e^{-i\omega t} \right], \quad (7.64)$$

Wave's partial reflection

and hence, according to Eq. (6),

$$H|_{z \leq 0} = \text{Re} \left[\frac{E_\omega}{Z_-(\omega)} \left(e^{ik_- z} - R e^{-ik_- z} \right) e^{-i\omega t} \right], \quad H|_{z \geq 0} = \text{Re} \left[\frac{E_\omega}{Z_+(\omega)} T e^{ik_+ z} e^{-i\omega t} \right]. \quad (7.65)$$

(Indices + and – correspond to, respectively, the media at $z > 0$ and $z < 0$.) Please note the following important features of these relations:

(i) Due to the problem linearity, we could (and did :-)) take the complex amplitudes of the reflected and transmitted wave proportional to that (E_ω) of the incident wave, describing them by the dimensionless coefficients R and T . The total reflection from an ideal mirror, that was discussed above, corresponds to the particular case $R = -1$ and $T = 0$.

(ii) Since the incident wave, that we are considering, arrives from one side only (from $z = -\infty$), there is no need to include a term proportional to $\exp\{-ik_+ z\}$ into Eqs. (64)-(65) - in our current problem. However, we would need such a term if the medium at $z > 0$ was non-uniform (e.g., had at least one more interface or any other inhomogeneity), because the wave reflected from that additional inhomogeneity would be incident on our interface (located at $z = 0$) from the right.

(iii) Solution (64)-(65) is sufficient even for the description of the cases when waves cannot propagate at $z \geq 0$, for example a conductor or a plasma with $\omega_p > \omega$. Indeed, the exponential drop of the field amplitude at $z > 0$ in such cases is automatically described by the imaginary part of wave number k_+ - see Eq. (29).

In order to find coefficients R and T , we need to use boundary conditions at $z = 0$. Since the reflection does not change the transverse character of the partial waves, at the normal incidence both vectors \mathbf{E} and \mathbf{H} remain tangential to the interface plane (in our notation, $z = 0$). Reviewing the arguments that has led us, in statics, to boundary conditions (3.47) and (5.118) for these components, we see that they remain valid for the time-dependent situation as well,²² so that for our current case of purely transverse waves we can write:

$$E|_{z=-0} = E|_{z=+0}, \quad H|_{z=-0} = H|_{z=+0}. \quad (7.66)$$

Plugging Eqs. (64)-(65) into these conditions, we get

$$1 + R = T, \quad \frac{1}{Z_-} (1 - R) = \frac{1}{Z_+} T. \quad (7.67)$$

²² For example, the first of conditions (66) may be obtained by integrating the full (time-dependent) Maxwell equation $\nabla \times \mathbf{E} + \partial \mathbf{B} / \partial t = 0$ over a narrow and long rectangular contour with dimensions l and d ($d \ll l$) stretched along the interface. In the Stokes theorem, the first term gives $\Delta E l$, which the contribution of the second term is proportional to product dl and vanishes as $d/l \rightarrow 0$. The proof of the second boundary condition is similar – as was already discussed in Sec. 6.2.

Solving this simple system of equations, we get²³

$$R = \frac{Z_+ - Z_-}{Z_+ + Z_-}, \quad T = \frac{2Z_+}{Z_+ + Z_-}. \quad (7.68)$$

These formulas are very important, and much more general than one may think, because they are applicable for virtually any 1D waves - electromagnetic or not, if only the impedance Z is defined in a proper way.²⁴ Since in the general case the wave impedances Z_{\pm} , defined by Eq. (28) with the corresponding indices, are complex functions of frequency, Eqs. (68) show that coefficients R and T may have imaginary parts as well. This fact has most important consequences at $z < 0$ where the reflected wave, proportional to R , interferes with the incident wave. Indeed, plugging $R = |R| e^{i\varphi}$ (where $\varphi \equiv \arg R$ is a real phase shift) into the expression in parentheses in the first of Eqs. (64), we may rewrite it as

$$\begin{aligned} (e^{ik_-z} + R e^{-ik_-z}) &= (1 - |R| + |R|)e^{ik_-z} + |R|e^{i\varphi} e^{-ik_-z} \\ &= (1 - |R|)e^{ik_-z} + 2|R|e^{i\varphi/2} \sin[k_-(z - \delta_-)], \quad \text{where } \delta_- \equiv \frac{\varphi - \pi}{2k_-}. \end{aligned} \quad (7.69)$$

This means that the field may be presented as a sum of a traveling wave and a standing wave, with amplitude proportional to $|R|$, shifted by distance δ_- toward the interface, relatively to the ideal-mirror pattern (61b). This effect is frequently used for the experimental measurements of an unknown impedance Z_+ of some medium, provided that Z_- is known (e.g., for the free space, $Z_- = Z_0$). For that, a small antenna (the *probe*), not disturbing the field distribution too much, is placed into the wave field, and the amplitude of the ac voltage induced in it by the wave in the probe is measured by some detector (e.g., a semiconductor diode with a quadratic I - V curve), as a function of z (Fig. 9). From this measurement, it is straightforward to find both $|R|$ and δ_- , and hence restore complex R , and then use Eq. (68) to calculate both modulus and argument of Z_+ .²⁵

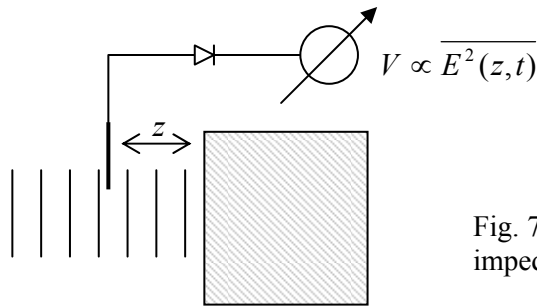


Fig. 7.9. Measurement of the complex impedance of a medium (schematically).

Now let us discuss what do these results give for waves incident from the free space ($Z(\omega) = Z_0 = \text{const}$, $k_- = k_0 = \omega/c$) onto the surface of two particular media.

²³ Please note that only the media impedances (rather than wave velocities) are important for the reflection in this case! Unfortunately, this fact is not clearly emphasized in some textbooks that discuss only the case $\mu_{\pm} = \mu_0$, when $Z = (\mu_0/\epsilon)^{1/2}$ and $v = 1/(\mu_0\epsilon)^{1/2}$ are proportional to each other.

²⁴ See, e.g., the discussion of elastic waves of mechanical deformations in CM Secs. 5.3, 5.4, 7.7, and 7.8.

²⁵ Before the advent of computers, specially lined paper (called the *Smith chart*) was commercially available for performing this recalculation graphically; it is occasionally used even nowadays for result presentation.

(i) For a collision-free plasma (with negligible magnetization) we may use Eq. (36) with $\mu(\omega) = \mu_0$, to present the impedance in either of two equivalent forms:

$$Z_+ = Z_0 \frac{\omega}{(\omega^2 - \omega_p^2)^{1/2}} = -iZ_0 \frac{\omega}{(\omega_p^2 - \omega^2)^{1/2}}. \quad (7.70)$$

The former expression is more convenient in the case $\omega > \omega_p$, when the wave vector k_+ and the wave impedance Z_+ of plasma are real, so that a part of the incident wave propagates into the plasma. Plugging this expression into the latter of Eqs. (68), we see that the transmission coefficient is real:

$$T = \frac{2\omega}{\omega + (\omega^2 - \omega_p^2)^{1/2}}. \quad (7.71)$$

Note that according to this formula, somewhat counter-intuitively, $T > 1$ for any frequency (above ω_p). How can the transmitted wave be more intensive than the incident one that has induced it? For a better understanding of this result, let us compare the powers (rather than amplitudes) of these two waves, i.e. their average Poynting vectors (42):

$$\overline{S}_{\text{incident}} = \frac{|E_\omega|^2}{2Z_0}, \quad \overline{S}_+ = \frac{|TE_\omega|^2}{2Z_+} = \frac{|E_\omega|^2}{2Z_0} \frac{4\omega(\omega^2 - \omega_p^2)^{1/2}}{[\omega + (\omega^2 - \omega_p^2)^{1/2}]^2}. \quad (7.72)$$

It is easy to see that the ratio of these two values²⁶ is always below 1 (and tends to zero at $\omega \rightarrow \omega_p$), so that only a fraction of the incident wave power may be transferred. Hence the result $T > 1$ may be interpreted as follows: the interface between two media also works as an *impedance transformer*: though it can never transfer more *power* than the incident wave provides, i.e. can only decrease the product $S = EH$, but since the ratio $Z = E/H$ changes at the interface, the amplitude of *one of the fields* may increase at the transfer.

Now let us proceed to case $\omega < \omega_p$, when the waves cannot propagate in the plasma. In this case, the latter of expressions (70) is more convenient, because it immediately shows that Z_+ is purely imaginary, while $Z_- = Z_0$ is purely real. This means that $(Z_+ - Z_-) = (Z_+ + Z_-)^*$, i.e. according to the first of Eqs. (68), $|R| = 1$, so that the reflection is total, i.e. no incident power (on the average) is transferred into the plasma – as was already discussed in Sec. 2. However, the complex R has a finite argument,

$$\varphi = \arg R = 2 \arg(Z_+ - Z_0) = -2 \arctan \frac{\omega}{(\omega_p^2 - \omega^2)^{1/2}}, \quad (7.73)$$

and hence provides a finite spatial shift (69) of the standing wave toward the plasma surface:

$$\delta_- = \frac{\varphi - \pi}{2k_0} = \frac{c}{\omega} \arctan \frac{\omega}{(\omega_p^2 - \omega^2)^{1/2}}. \quad (7.74)$$

On the other hand, we already know from Eq. (40) that the solution at $z > 0$ is exponential, with the decay length δ that is described by Eq. (39). Calculating, from coefficient T , the exact coefficient before this exponent, it is straightforward to verify that the electric and magnetic fields are indeed

²⁶ This ratio is sometimes also called the transmission coefficient, but in order to avoid its confusion with T , it is better to call it the *power transmission coefficient*.

continuous at the interface, forming the pattern shown by red lines in Fig. 8. This penetration may be experimentally observed, for example, by bringing close to the interface the surface of another material transparent as frequency ω . Even without solving this problem exactly, it is evident that if the distance between these two interfaces becomes comparable to δ , a part of the exponential “tail” of the field is picked up by the second material, and induces a propagating wave. This is an electromagnetic analog of the quantum-mechanical tunneling through a potential barrier.²⁷

Note that at $\omega \ll \omega_p$, both \mathcal{S} and δ are reduced to the same frequency-independent value,

$$\delta, \delta_- \rightarrow \frac{c}{\omega_p} = \left(\frac{c^2 \epsilon_0 m_e}{n e^2} \right)^{1/2} = \left(\frac{m_e}{\mu_0 n e^2} \right)^{1/2}, \quad (7.75)$$

which is just the field penetration depth δ (6.38) calculated for a perfect conductor model (assuming $m = m_e$ and $\mu = \mu_0$) in the quasistatic limit. This is natural, because the condition $\omega \ll \omega_p$ may be recast as $\lambda_0 = 2\pi c/\omega \gg 2\pi c/\omega_p = 2\pi\delta$.

(ii) Now let us consider electromagnetic wave reflection from a nonmagnetic conductor. In the simplest low-frequency limit, when $\omega\tau$ is much less than 1, the conductor may be described by a frequency-independent conductivity σ .²⁸ According to Eq. (46), in this case we can take

$$Z_+ = \left(\frac{\mu_0}{\epsilon_{\text{opt}}(\omega) + i\sigma/\omega} \right)^{1/2}. \quad (7.76)$$

With this substitution, Eqs. (68) immediately give us all the results of interest. In particular, they show that now R is complex, and hence some fraction F of the incident wave is absorbed by the conductor. Using Eq. (42), we may calculate the fraction to be

$$F \equiv \frac{\overline{S_+}|_{z=+0}}{S_{\text{incident}}} = |T|^2 \text{Re} \frac{Z_0}{Z_+}. \quad (7.77)$$

(Since power flow S_+ into the conductor depends on z , tending to zero at distances $z \sim \delta$, it is important to calculate it directly at the interface to account for the absorption in the whole volume of the conductor.) Restricting ourselves, for the sake of simplicity, to the most important quasistatic limit, i.e. to $Z_+ = (\mu_0 \omega / i \sigma)^{1/2}$, and using Eq. (6.27) to express the impedance via the skin depth, $Z_+ = \pi(2/i)^{1/2}(\delta_s/\lambda_0)Z_0$, we see that $|Z_+| \ll Z_0$, so that, according to Eq. (68), $T \approx 2Z_+/Z_0$ and

Wave's
absorption
in
conductor's
surface

$$F \approx \frac{4|Z_+|^2}{Z_0^2} \text{Re} \frac{Z_0}{Z_+} = 2 \frac{\delta_s}{\lambda_0} \ll 1. \quad (7.78)$$

Thus the absorbed power scales as the ratio of the skin depth to the free-space wavelength. This important result is widely used for the semi-qualitative evaluation of power losses in metallic waveguides and resonators, and immediately shows that in order to keep the losses low, the characteristic size of such systems (that gives a scale of the free-space wavelengths λ_0 , at which they are

²⁷ See, e.g., QM Sec. 2.3.

²⁸ In a typical metal, $\tau \sim 10^{-13}$ s, so that this approximation work well all the way up to $\omega \sim 10^{13}$ s⁻¹, i.e. up to the far-infrared frequencies.

used) should be much larger than δ_s . A more detailed theory of these structures will be discussed later in this chapter.

7.5. Refraction

Now let us consider the effects arising at the plane interface if the wave incidence angle θ (Fig. 10) is arbitrary, rather than equal to zero as in our previous analysis, for the simplest case of fully transparent media, with real ε_{\pm} and μ_{\pm} .

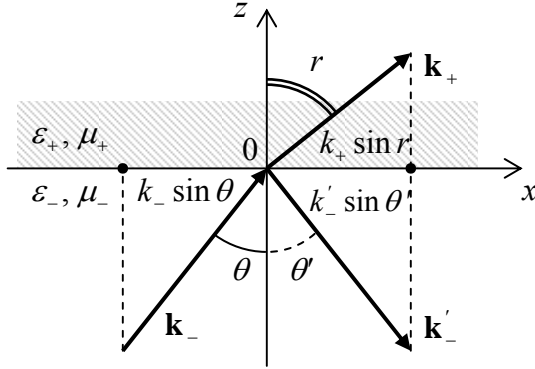


Fig. 7.10. Plane wave reflection, transmission, and refraction at a plane interface. The plane of drawing is selected to contain all three wave vectors \mathbf{k}_+ , \mathbf{k}_- , and \mathbf{k}'_- .

In contrast with the case of normal incidence, here the wave vectors \mathbf{k}_- , \mathbf{k}'_- , and \mathbf{k}_+ of the three component (incident, reflected, and transmitted) waves may have different directions. Hence now we have to start our analysis with writing a general expression for a single plane, monochromatic wave for the case when its wave vector \mathbf{k} has all 3 Cartesian components, rather than one. An evident generalization of Eq. (11) to this case is

$$f(\mathbf{r}, t) = \text{Re} \left[f_{\omega} e^{i(k_x x + k_y y + k_z z) - \omega t} \right] = \text{Re} \left[f_{\omega} e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} \right]. \quad (7.79)$$

This relation enables a ready analysis of “kinematic” relations that are independent of the media impedances. Indeed, it is sufficient to notice that in order to satisfy *any* linear, homogeneous boundary conditions at the interface ($z = 0$), all waves have the same temporal and spatial dependence on this plane. Hence if we select plane xz so that vector \mathbf{k}_- lies in it, then $(k_-)_y = 0$, and \mathbf{k}_+ and \mathbf{k}'_- cannot have any y -component either, i.e. all three vectors lie in the same plane - that is selected as the plane of drawing of Fig. 10. Moreover, due to the same reason their x -components should be equal:

$$k_- \sin \theta = k'_- \sin \theta' = k_+ \sin r. \quad (7.80)$$

From here we immediately have the well-known laws of reflection

$$\theta' = \theta,$$

$$(7.81) \quad \text{Reflection angle}$$

and refraction:²⁹

²⁹ This relation is traditionally called the *Snell law*, after a 17th century's author W. Snellius, though it has been traced back to a circa 984 manuscript by Abu Saad al-Ala ibn Sahl.

Snell
law

$$\frac{\sin r}{\sin \theta} = \frac{k_-}{k_+}. \quad (7.82)$$

In this form, the laws are valid for plane waves of any nature. In optics, the Snell law (82) is frequently presented in the form

$$\frac{\sin r}{\sin \theta} = \frac{n_-}{n_+}, \quad (7.83)$$

where n_{\pm} is the *index of refraction* (also called the “refractive index”) of the corresponding medium, defined as its wave number normalized so that of the free space (at wave’s frequency):

Index
of refraction

$$n_{\pm} \equiv \frac{k_{\pm}}{k_0} = \left(\frac{\epsilon_{\pm} \mu_{\pm}}{\epsilon_0 \mu_0} \right)^{1/2}. \quad (7.84)$$

Perhaps the most famous corollary of the Snell law is that if a wave propagates from a medium with a higher index of refraction to that with a lower one (i.e. if $n_- > n_+$ in Fig. 10), for example from water into air, there is always a certain *critical* value θ_c of the incidence angle,

Critical
angle

$$\theta_c = \arcsin \frac{n_+}{n_-} = \arcsin \left(\frac{\epsilon_+ \mu_+}{\epsilon_- \mu_-} \right)^{1/2}, \quad (7.85)$$

at which angle r reaches $\pi/2$. At a larger θ , i.e. within the range $\theta_c < \theta < \pi/2$, the boundary conditions cannot be satisfied with a refracted wave with a real wave vector, so that the wave experiences the so-called *total internal reflection*. This effect is very important for practice, because it shows that dielectric surfaces may be used as mirrors, in particular in optical fibers - to be discussed in more detail in Sec. 8 below. This is very fortunate for all the telecommunication technology, because the light reflection from metals is rather imperfect. Indeed, according to Eq. (78), in the optical range ($\lambda_0 \sim 0.5 \mu\text{m}$, i.e. $\omega \sim 10^{15} \text{s}^{-1}$), even the best conductors (with $\sigma \sim 6 \times 10^8 \text{S/m}$ and hence the normal skin depth $\delta_s \sim 1.5 \text{nm}$) provide relatively high losses $F \sim 1\%$ at each reflection.

Note, however, that even within the range $\theta_c < \theta < \pi/2$ the field at $z > 0$ is not identically equal to zero: just as it does at the normal incidence ($\theta = 0$), it penetrates into the less dense media by a distance of the order of λ_0 , exponentially decaying inside it. At $\theta \neq 0$ the penetrating field still changes sinusoidally, with wave number (80), along the interface. Such a field, exponentially dropping in one direction but still propagating as a wave in another direction, is frequently called the *evanescent wave*.

One more remark: just as at the normal incidence, the field penetration into another medium causes a phase shift of the reflected wave – see, e.g., Eq. (69) and its discussion. A new feature of this phase shift, arising at $\theta \neq 0$, is that it also has a component parallel to the interface – the so-called *Goos-Hänchen effect*. In geometric optics, this effect leads to an image shift (relative to that its position in a perfect mirror) with components both normal and parallel to the interface.

Now let us carry out an analysis of the “dynamic” relations that determine amplitudes of the refracted and reflected waves. For this we need to write explicitly the boundary conditions at the interface (i.e. plane $z = 0$). Since now the electric and/or magnetic fields may have components normal to the plane, in addition to the continuity of their tangential components, which we have repeatedly discussed,

$$E_{x,y}|_{z=-0} = E_{x,y}|_{z=+0}, \quad H_{x,y}|_{z=-0} = H_{x,y}|_{z=+0}, \quad (7.86)$$

we also need relations for the normal components. As it follows from the homogeneous macroscopic Maxwell equations (6.94b), they are also the same as in statics ($D_n = \text{const}$, $B_n = \text{const}$), for our reference frame choice (Fig. 10) giving

$$\varepsilon_- E_z|_{z=-0} = \varepsilon_+ E_z|_{z=+0}, \quad \mu_- H_z|_{z=-0} = \mu_+ H_z|_{z=+0}. \quad (7.87)$$

The expressions of these components via amplitudes E_ω , RE_ω , and TE_ω of the incident, reflected and transmitted waves depend on the incident wave's polarization. For example, for a linearly-polarized wave with the electric field vector *perpendicular* to the plane of incidence (Fig. 11a), i.e. *parallel* to the interface plane, the reflected and refracted waves are similarly polarized.

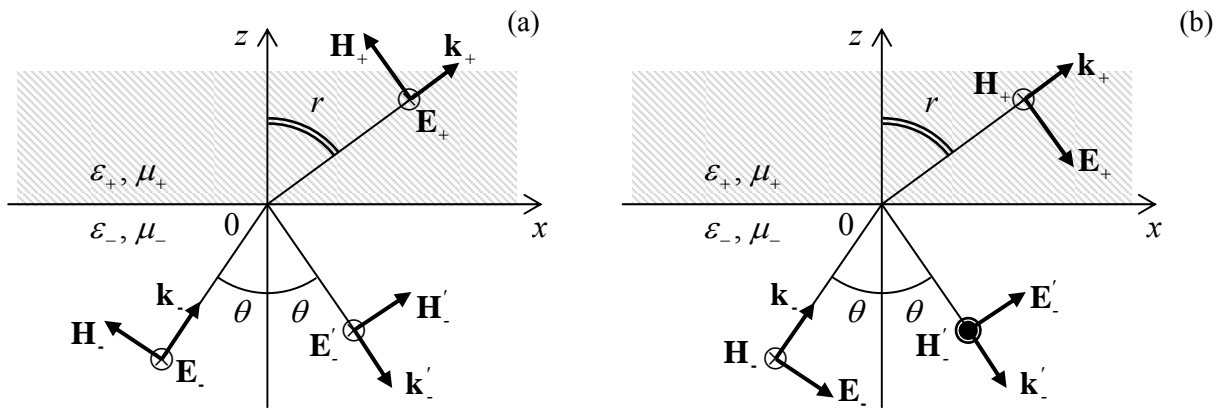


Fig. 7.11. Reflection and refraction at two different linear polarizations of the incident wave.

As a result, all E_z are equal to zero (so that the first of Eqs. (87) is inconsequential), while the tangential components of the electric field are just equal to their full amplitudes, just as at the normal incidence, so we still can use Eqs. (64) to express these components via coefficients R and T . However, at $\theta \neq 0$ the magnetic fields have not only tangential components

$$H_x|_{z=-0} = \text{Re} \left[\frac{E_\omega}{Z_-} (1 - R) \cos \theta e^{-i\omega t} \right], \quad H_x|_{z=+0} = \text{Re} \left[\frac{E_\omega}{Z_+} T \cos r e^{-i\omega t} \right], \quad (7.88)$$

but also normal components (Fig. 11a):

$$H_z|_{z=-0} = \text{Re} \left[\frac{E_\omega}{Z_-} (1 + R) \sin \theta e^{-i\omega t} \right], \quad H_z|_{z=+0} = \text{Re} \left[\frac{E_\omega}{Z_+} T \sin r e^{-i\omega t} \right]. \quad (7.89)$$

Plugging these expressions into the boundary conditions expressed by Eqs. (86) (in this case, for y components only) and the second of Eqs. (87), we get *three* equations for *two* unknown coefficients R and T . However, two of these equations duplicate each other because of the Snell law, and we get just two independent equations,

$$1 + R = T, \quad \frac{1}{Z_-}(1 - R)\cos\theta = \frac{1}{Z_+}T\cos r, \quad (7.90)$$

which are a very natural generalization of Eqs. (67), with replacements $Z_- \rightarrow Z_- \cos r$, $Z_+ \rightarrow Z_+ \cos\theta$. As a result, we can immediately use Eq. (68) to write the solution of system (90):³⁰

$$R = \frac{Z_+ \cos\theta - Z_- \cos r}{Z_+ \cos\theta + Z_- \cos r}, \quad T = \frac{2Z_+ \cos\theta}{Z_+ \cos\theta + Z_- \cos r}. \quad (7.91a)$$

If we want to express the coefficients via the angle of incidence alone, we should use the Snell law (82) to eliminate angle r , getting

$$R = \frac{Z_+ \cos\theta - Z_- [1 - (k_- / k_+)^2 \sin^2 \theta]^{1/2}}{Z_+ \cos\theta + Z_- [1 - (k_- / k_+)^2 \sin^2 \theta]^{1/2}}, \quad T = \frac{2Z_+ \cos\theta}{Z_+ \cos\theta + Z_- [1 - (k_- / k_+)^2 \sin^2 \theta]^{1/2}}. \quad (7.91b)$$

However, my strong preference is to use the kinematic relation (82) and dynamic relations (91a) separately, because Eq. (91b) obscures the very important physical fact that the ratio of k_{\pm} , i.e. of the wave velocities of the two media, is only involved in the Snell law (79), while the dynamic relations essentially include only the ratio of wave impedances - just as in the case of normal incidence.

In the opposite case of the linear polarization of the electric field within the plane of incidence (Fig. 11b), it is the magnetic field that does not have a normal component, so it is now the second of Eqs. (87) that does not participate in the solution. However, now the electric fields in two media have not only tangential components,

$$E_x|_{z=0} = \text{Re}[E_{\omega}(1 + R)\cos\theta e^{-i\omega t}], \quad E_x|_{z=0} = \text{Re}[E_{\omega}T\cos r e^{-i\omega t}] \quad (7.92)$$

but also normal components (Fig. 11b):

$$E_z|_{z=0} = E_{\omega}(-1 + R)\sin\theta, \quad E_z|_{z=0} = -E_{\omega}T\sin r. \quad (7.93)$$

As a result, instead of Eqs. (90), the reflection and transmission coefficients are related as

$$(1 + R)\cos\theta = T\cos r, \quad \frac{1}{Z_-}(1 - R) = \frac{1}{Z_+}T. \quad (7.94)$$

Again, the solution of this system may be immediately written using the analogy with Eq. (67):

$$R = \frac{Z_+ \cos r - Z_- \cos\theta}{Z_+ \cos r + Z_- \cos\theta}, \quad T = \frac{2Z_+ \cos\theta}{Z_+ \cos r + Z_- \cos\theta}, \quad (7.95a)$$

or, alternatively, using the Snell law:

$$R = \frac{Z_+ [1 - (k_- / k_+)^2 \sin^2 \theta]^{1/2} - Z_- \cos\theta}{Z_+ [1 - (k_- / k_+)^2 \sin^2 \theta]^{1/2} + Z_- \cos\theta}, \quad T = \frac{2Z_+ \cos\theta}{Z_+ [1 - (k_- / k_+)^2 \sin^2 \theta]^{1/2} + Z_- \cos\theta}. \quad (7.95b)$$

³⁰ Note that we may calculate the reflection and transmission coefficients R' and T' for the wave traveling in the opposite direction just by making parameter swaps $Z_+ \leftrightarrow Z_-$ and $\theta \leftrightarrow r$, and that the resulting coefficients satisfy the following *Stokes relations*: $R' = -R$, and $R^2 + TT' = 1$, for any Z_{\pm} .

For the particular case $\mu_+ = \mu_- = \mu_0$, when $Z_+/Z_- = (\varepsilon_+/\varepsilon_-)^{1/2} = k_-/k_+ = n_-/n_+$ (which is approximately correct for traditional optical media), Eqs. (91b) and (95b) are called the *Fresnel formulas*.³¹ Most textbooks are quick to point out that there is a major difference between these cases: while for the electric field polarization within the plane of incidence (Fig. 11b), the reflected wave amplitude (proportional to coefficient R) turns to zero at a special value of θ (the so-called *Brewster angle*):³²

$$\theta_B = \arctan \frac{n_+}{n_-}, \quad (7.96)$$

while there is no such angle in the opposite case (Fig. 11a).³³ However, that this statement, as well as Eq. (96), is true only for the case $\mu_+ = \mu_-$. In the general case of different ε and μ , Eqs. (91) and (95) show that the reflected wave vanishes at $\theta = \theta_B$ with

$$\tan^2 \theta_B = \frac{\varepsilon_- \mu_+ - \varepsilon_+ \mu_-}{\varepsilon_+ \mu_+ - \varepsilon_- \mu_-} \times \begin{cases} (\mu_+ / \mu_-), & \text{for } \mathbf{E} \perp \mathbf{n}_z \text{ (Fig. 11a),} \\ (-\varepsilon_+ / \varepsilon_-), & \text{for } \mathbf{H} \perp \mathbf{n}_z \text{ (Fig. 11b).} \end{cases} \quad (7.97) \quad \text{Brewster angle}$$

Note the natural $\varepsilon \leftrightarrow \mu$ symmetry of these relations, resulting from the $\mathbf{E} \leftrightarrow \mathbf{H}$ symmetry for these two polarization cases (Fig. 11). They also show that for any set of parameters of the two media (with $\varepsilon_{\pm}, \mu_{\pm} > 0$), $\tan^2 \theta_B$ is positive (and hence a real Brewster angle θ_B exists) only for one of these two polarizations. In particular, if the interface is due to the change of μ alone (i.e. $\varepsilon_+ = \varepsilon_-$), the first of Eqs. (97) is reduced to the simple form (96) again, while for the polarization shown in Fig. 11b there is no Brewster angle, i.e. the reflected wave has a nonvanishing amplitude for any θ .

Such account of both media parameters on an equal footing is especially necessary to describe the so-called *negative refraction* effects.³⁴ As was shown in Sec. 2, in a medium with electric-field-driven resonances, function $\varepsilon(\omega)$ may be almost real and negative, at least within limited frequency intervals – see, in particular, Eq. (34) and Fig. 5. As have already been discussed, if, at these frequencies, function $\mu(\omega)$ is real and positive, then $k^2(\omega) = \omega^2 \varepsilon(\omega) \mu(\omega) < 0$, and k may be presented as i/δ with real δ , meaning the exponential field decay into the medium. However, let consider the case when both $\varepsilon(\omega) < 0$ and $\mu(\omega) < 0$ at a certain frequency. (This is evidently possible in a medium with both \mathbf{E} -driven and \mathbf{H} -driven resonances, at proper relations between their eigenfrequencies.) Since in this case $k^2(\omega) = \omega^2 \varepsilon(\omega) \mu(\omega) > 0$, the wave vector is real, so that Eq. (79) describes a traveling wave, and one could think that there is nothing new in this case. Not quite so!

³¹ After A.-J. Fresnel (1788-1827), one of the pioneers of the wave optics, who is credited, among many other contributions (see in particular Ch. 8), for the concept of light as a purely transverse wave.

³² A very simple interpretation of Eq. (93) is based on the fact that, together with the Snell law (82), it gives $r + \theta = \pi/2$. As a result, vector \mathbf{E}_+ is parallel to vector \mathbf{k}_- , and hence oscillating dipoles of medium at $z > 0$ do not have the component which could induce the transverse electric field \mathbf{E}_- of the reflected wave.

³³ This effect is used in practice to obtain linearly polarized light, with the electric field vector perpendicular to the plane of incidence, from the natural light with its random polarization. An even more practical application of the effect is a partial reduction of undesirable glare from wet surfaces (for the water/air interface, $n_+/n_- \approx 1.33$, giving $\theta_B \approx 50^\circ$) by making car light covers and sunglasses of vertically-polarizing materials.

³⁴ Despite some important background theoretical work by A. Schuster (1904), L. Mandelstam (1945), D. Sivikhin (1957), and especially V. Veselago (1966-67), the negative refractivity effects have only recently become a subject of intensive scientific research and engineering development.

First of all, for a sinusoidal, plane wave (79), operator ∇ is equivalent to the multiplication by $i\mathbf{k}$. As the Maxwell equations (2a) show, this means that at a fixed direction of vectors \mathbf{E} and \mathbf{k} , the simultaneous reversal of signs of ε and μ means the reversal of the direction of vector \mathbf{H} . Namely, if both ε and μ are positive, these equations are satisfied with mutually orthogonal vectors \mathbf{E} , \mathbf{H} , and \mathbf{k} forming the usual, *right-hand* system (see Fig. 1 and Fig. 12a), the name stemming from the popular “right-hand rule” used to determine the vector product direction. However, if both ε and μ are negative, the vectors form a *left-hand* system – see Fig. 12b. (Due to this fact, the media with $\varepsilon < 0$ and $\mu < 0$ are frequently called the *left-handed materials*, LHM for short.) According to Eq. (6.97), that does not involve media parameters, this means that for a plane wave in a left-hand material, the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$, i.e. of the energy flow, is directed *opposite* to the wave vector \mathbf{k} .

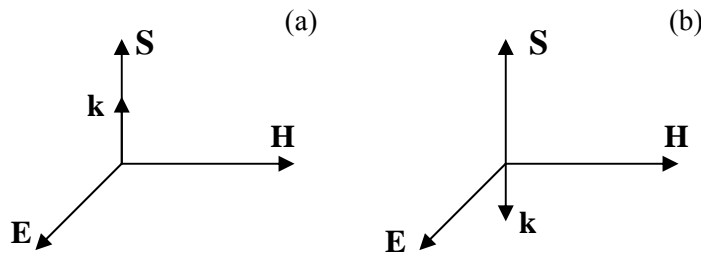


Fig. 7.12. Directions of main vectors of a plane wave inside a medium with (a) positive and (b) negative ε and μ .

This fact may seem strange, but is in no contradiction with any fundamental principle. Let me remind you that, according to the definition of vector \mathbf{k} , its direction shows the direction of the *phase* velocity $v_{ph} = \omega/k$ of a sinusoidal (and hence infinitely long) wave that cannot be used, for example, for signaling. Such signaling (by sending wave packets – see Fig. 13) is possible with the *group* velocity $v_{gr} = d\omega/dk$. This velocity in left-hand materials is always positive (directed along vector \mathbf{S}).

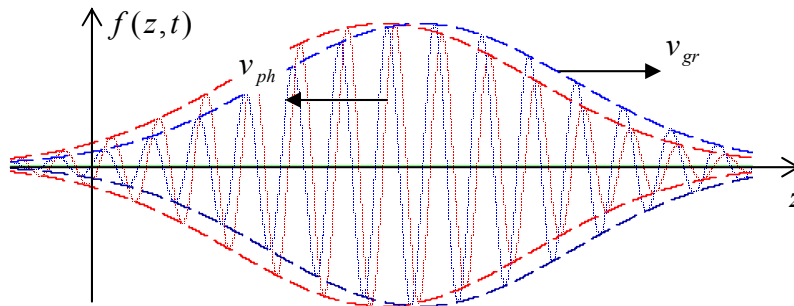


Fig. 7.13. Example of a wave packet moving along axis z with a negative phase velocity, but positive group velocity. Blue lines show a packet snapshot a short time interval after the first snapshot (red lines).

Maybe the most fascinating effect possible with left-hand materials is the wave refraction at their interfaces with the usual, right-handed materials - first predicted by V. Veselago. Consider the example shown in Fig. 14a. In the incident wave, coming from the usual material, the directions of vectors \mathbf{k} and \mathbf{S} coincide, and so they are in the reflected wave characterized by vectors \mathbf{k}' and \mathbf{S}' . This means that the electric and magnetic fields in the interface plane ($z = 0$) are, at our choice of coordinates, proportional to $\exp\{ik_x x\}$, with positive component $k_x = k \cos \theta$. In order to satisfy any linear boundary conditions, the refracted wave, going into the left-handed material, should match that dependence, i.e.

have a positive x -component of its wave vector \mathbf{k}_+ . But in this medium, this vector has to be antiparallel to vector \mathbf{S} that, in turn, should be directed out of the interface, because it presents the power flow from the interface into the material bulk. These conditions cannot be reconciled by the refracted wave propagating along the usual Snell-law direction (shown by the dashed line in Fig. 13a), but are all satisfied at refraction in the direction given by Snell's angle with negative sign. (Hence the term “negative refraction”).³⁵

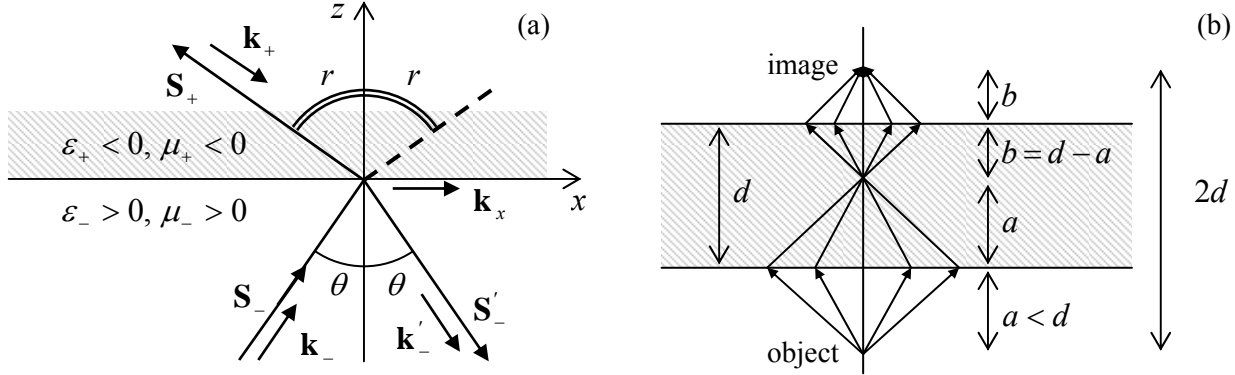


Fig. 7.14. Negative refraction: (a) waves at the interface between media with positive and negative values of $\epsilon\mu$, and (b) the hypothetical *perfect lense*: a parallel plate made of a material with $\epsilon = -\epsilon_0$ and $\mu = -\mu_0$.

In order to understand how unusual the results of the negative refraction may be, let us consider a parallel slab of thickness d , made of a hypothetical left-handed material with $\epsilon = -\epsilon_0$, $\mu = -\mu_0$ (Fig. 14b), placed in free space. For such a material, the refraction angle $r = -\theta$, so that the rays from a point source, located at a distance $a < d$ from the slab, propagate as shown in that figure, i.e. all meet again at distance a inside the plate, and then continue to propagate to the second surface of the slab. Repeating our discussion for this surface, we see that a point's image is also formed beyond the plate at distance $2a + 2b = 2a + 2(d - a) = 2d$ from the object. Superficially, this looks like the usual lense, but the well-known lense formula, which relates a and b with the focal length f , is *not* satisfied. (In particular, a parallel beam is *not* focused into a point at any finite distance.)

As an additional difference from the usual lense, the system shown in Fig. 14b *does not reflect* any part of the incident light. Indeed, it is straightforward to check that in order for all above formulas for R and T to be valid, the sign of the wave impedance Z in left-handed materials has to be kept positive. Thus, for our particular choice of parameters ($\epsilon = -\epsilon_0$, $\mu = -\mu_0$), Eqs. (91a) and (95a) are valid with $Z_+ = Z_- = Z_0$ and $\cos r = \cos \theta = 1$, giving $R = 0$ for any linear polarization, and hence for any other wave polarization - circular, elliptic, natural, etc.

The perfect lense suggestion has triggered a wave of efforts to implement left-hand materials experimentally. (Attempts to found such materials in nature have failed so far.) Most progress in this direction has been achieved using the so-called *metamaterials*, which are essentially quasi-periodic arrays of specially designed electromagnetic resonators, ideally with high density $n \gg \lambda^{-3}$. For example,

³⁵ Inspired by this fact, in some publications the left-hand materials are prescribed a negative index of refraction n . However, this prescription should be treated with care (for example, it complies with the first form of Eq. (84), but not its second form), and the sign of n , in contrast to that of wave vector \mathbf{k} , is the matter of convention.

Fig. 15a shows the metamaterial that was used for the first demonstration of negative refractivity in the microwave region, i.e. a few-GHz frequencies – see Fig. 15b. It combines straight strips of a metallic film, working as lumped resonators with a large electric dipole moment (hence strongly coupled to wave's electric field \mathbf{E}), and several almost-closed film loops (so-called *split rings*), working as lumped resonators with large magnetic dipole moments, coupled to field \mathbf{H} . By designing the resonance frequencies close to each other, the negative refractivity may be achieved – see the black line in Fig. 15b, which shows experimental data. Recently, the negative refractivity was demonstrated in the optical range, albeit at relatively large absorption that spoils all potentially useful features of the left-handed materials.

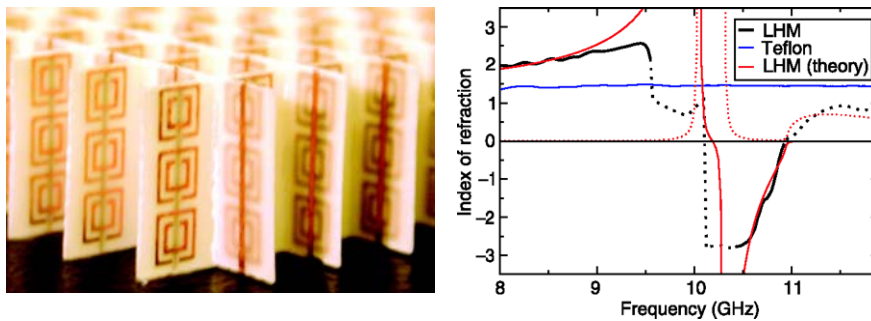


Fig. 7.15. The first artificial left-hand material with experimentally demonstrated negative refraction in a microwave region. Adapted from R. Shelby *et al.*, *Science* **292**, 77 (2001). © AAAS.

This progress has stimulated the development of other potential uses of metamaterials (not necessarily the left-handed ones), in particular designs of nonuniform systems with engineered distributions $\epsilon(\mathbf{r}, \omega)$ and $\mu(\mathbf{r}, \omega)$, which may provide electromagnetic wave propagation along the desired paths, e.g. around a certain region of space (Fig. 16), making it virtually invisible for an external observer - so far, within a limited frequency range, and a certain wave polarization only. Due to these restrictions, the practical value of this work on such *invisibility cloaks* is not yet clear (at least to this author); but so much attention is focused on this issue³⁶ that the situation should become much more clear in just a few years.

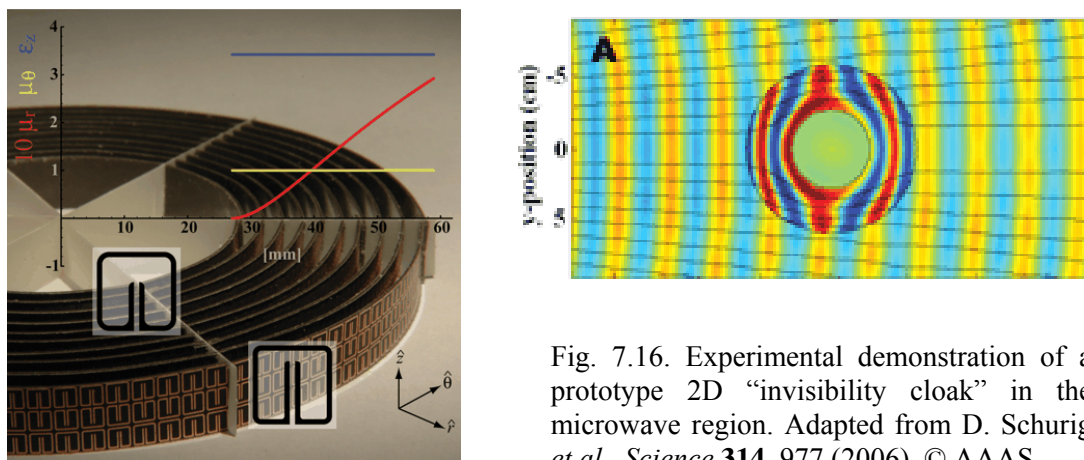


Fig. 7.16. Experimental demonstration of a prototype 2D “invisibility cloak” in the microwave region. Adapted from D. Schurig *et al.*, *Science* **314**, 977 (2006). © AAAS.

³⁶ For a recent review, see, e.g., B. Wood, *Comptes Rendus Physique* **10**, 379 (2009).

7.6. Transmission lines: TEM waves

So far, we have analyzed plane the electromagnetic waves with infinite cross-section. The cross-section may be limited, still sustaining wave propagation, using *wave transmission lines* (also called *waveguides*): cylindrically-shaped structures made of either good conductors or dielectrics. Let us first discuss the first option. In order to keep our analysis (relatively :-) simple, let us assume that:

(i) the structure is a cylinder (not necessarily with a round cross-section, see Fig. 17) filled with a usual (right-handed), uniform dielectric material with negligible losses: $\varepsilon = \varepsilon' > 0$, $\mu = \mu' > 0$, and

(ii) the wave attenuation due to the skin effect is also negligibly low. (As Eq. (78) indicates, for that the characteristic size a of waveguide's cross-section has to be much larger than the skin-depth δ_s of its wall material. The effect of skin-effect losses will be analyzed in Sec. 10 below.)

After such exclusion of attenuation, we may look for a particular solution of the Maxwell equations in the form of a monochromatic wave traveling along the waveguide:

$$\mathbf{E}(\mathbf{r}, t) = \text{Re}[\mathbf{E}_\omega(x, y)e^{i(k_z z - \omega t)}] \quad \mathbf{H}(\mathbf{r}, t) = \text{Re}[\mathbf{H}_\omega(x, y)e^{i(k_z z - \omega t)}], \quad (7.98)$$

with real k_z . Note that this form allows an account for a substantial coordinate dependence of the electric and magnetic field in the plane $\{x, y\}$ of the waveguide's cross-section, as well as for longitudinal components of the fields, so that solution (98) is substantially more complex than the plane waves we have discussed above. We will see in a minute that as a result of this dependence, constant k_z may be very much different from the plane-wave value (13), $k \equiv \omega(\varepsilon\mu)^{1/2}$, in the same material.

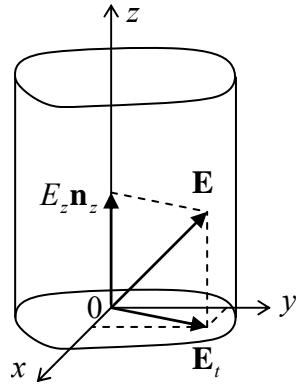


Fig. 7.17. Decomposition of the electric field in a waveguide.

In order to describe these effects explicitly, let us decompose the complex amplitudes of the fields into the longitudinal and transverse components (Fig. 17)³⁷

$$\mathbf{E}_\omega = E_z \mathbf{n}_z + \mathbf{E}_t, \quad \mathbf{H}_\omega = H_z \mathbf{n}_z + \mathbf{H}_t. \quad (7.99)$$

Plugging Eqs. (98)-(99) into the homogeneous Maxwell equations (2), and requiring the longitudinal and transverse components to be balanced separately, we get

³⁷ Note that for the notation simplicity, I am dropping index ω in the complex amplitudes of the field components, and later will drop argument ω in k_z and Z , though they may depend on the wave frequency rather substantially – see below.

$$\begin{aligned}
ik_z \mathbf{n}_z \times \mathbf{E}_t - i\omega\mu \mathbf{H}_t &= -\nabla_t \times (E_z \mathbf{n}_z), & ik_z \mathbf{n}_z \times \mathbf{H}_t + i\omega\varepsilon \mathbf{E}_t &= -\nabla_t \times (H_z \mathbf{n}_z), \\
\nabla_t \times \mathbf{E}_t &= i\omega\mu H_z \mathbf{n}_z, & \nabla_t \times \mathbf{H}_t &= -i\omega\varepsilon E_z \mathbf{n}_z, \\
\nabla_t \cdot \mathbf{E}_t &= -ik_z E_z, & \nabla_t \cdot \mathbf{H}_t &= -ik_z H_z.
\end{aligned} \tag{7.100}$$

where ∇_t is the 2D Laplace operator acting in the transverse plane $[x, y]$. These equations may look even more bulky than the original Maxwell equations, but actually are much simpler for analysis. Indeed, eliminating the transverse components from these equations (or, even simpler, just plugging Eq. (99) into Eqs. (3) and keeping just their z -components), we may get a pair of self-consistent equations for the longitudinal components of the fields,³⁸

2D Helmholtz
equations for
 E_z and H_z

$$(\nabla_t^2 + k^2) E_z = 0, \quad (\nabla_t^2 + k^2) H_z = 0, \tag{7.101}$$

where k is still defined by Eq. (13), $k = (\varepsilon\mu)^{1/2} \omega$, and

Wave vector
component
balance

$$k_t^2 \equiv k^2 - k_z^2 = \omega^2 \varepsilon\mu - k_z^2. \tag{7.102}$$

After distributions $E_z(x, y)$ and $H_z(x, y)$ have been found from these equations, they provide right-hand parts for rather simple, closed system of equations (100) for the transverse components of field vectors. Moreover, as we will see below, each of the following three types of solutions:

- (i) with $E_z = 0$ and $H_z = 0$ (called the *transverse*, or *TEM waves*),
- (ii) with $E_z = 0$, but $H_z \neq 0$ (called either *TE waves* or, more frequently, *H modes*), and
- (iii) with $E_z \neq 0$, but $H_z = 0$ (*TM waves* or *E modes*),

has its own dispersion law and hence wave propagation velocity; as a result, these *modes* (the term meaning the field distribution pattern) may be considered separately.

Let us start with the simplest, TEM waves with no longitudinal components of either field. For them, the top two equations of system (100) immediately give Eqs. (6) and (13), and $k_z = k$. In plain English, this means that $\mathbf{E} = \mathbf{E}_t$ and $\mathbf{H} = \mathbf{H}_t$ are proportional to each other and mutually perpendicular (just as in the plane wave) at each point of the cross-section, and that the TEM wave impedance $Z \equiv E/H$ and dispersion law $\omega(k)$, and hence the propagation speed, are the same as in a plane wave in the material filling the waveguide. In particular, if ε and μ are frequency-independent within a certain frequency range, the dispersion law is linear, $\omega = k/(\varepsilon\mu)^{1/2}$, and wave's speed does not depend on its frequency. For practical applications to telecommunications, this is a very important advantage of TEM waves over their TM and TE counterparts – to be discussed below.

Unfortunately, such waves cannot propagate in every waveguide. In order to show this, let us have a look at the two last lines of Eqs. (100). For the TEM waves ($E_z = 0$, $H_z = 0$, $k_z = k$), they yield

$$\begin{aligned}
\nabla_t \times \mathbf{E}_t &= 0, & \nabla_t \times \mathbf{H}_t &= 0, \\
\nabla_t \cdot \mathbf{E}_t &= 0, & \nabla_t \cdot \mathbf{H}_t &= 0.
\end{aligned} \tag{7.103}$$

In the macroscopic approximation of the boundary conditions (i. e., neglecting the screening and skin depths), we have to require that the wave does not penetrate the walls, so that inside them, $\mathbf{E} = \mathbf{H} = 0$. Close to the wall but inside the waveguide, the normal component E_n of the electric field may be

³⁸ The wave equation presented in the form (98) is called the (in our particular case, 2D) *Helmholtz equation*, after H. von Helmholtz (1821-1894) - the mentor of H. Hertz and M. Planck, among many others.

different from zero, because surface charges may sustain its jump (see Sec. 2.1). Similarly, the tangential component H_τ of the magnetic field may have a finite jump at the surface due to skin currents. However, the tangential component of the electric field and the normal component of magnetic field cannot experience such jump, and in order to have them vanishing inside the walls they have to equal zero near the walls inside the waveguide as well:

$$\mathbf{E}_\tau = 0, \quad H_n = 0. \quad (7.104)$$

But the left columns of Eqs. (103) and (104) coincide with the formulation of the 2D boundary problem of electrostatics for the electric field induced by electric charges of the conducting walls, with the only difference that in our current case the value of ε should be replaced with $\varepsilon(\omega)$. Similarly, the right columns of those relations coincide with the formulation of the 2D boundary problem of magnetostatics for the magnetic field induced by currents in the walls, with $\mu = \mu(\omega)$. The only difference is that in our current case the magnetic fields should not penetrate inside the conductors.

Now we immediately see that in waveguides with a singly-connected wall topology (see, e.g., the particular example shown in Fig. 17), TEM waves are impossible, because there is no way to create a finite electrostatic field inside a conductor with such cross-section. Fortunately, such fields (and hence TEM waves) are possible in structures with cross-sections consisting of two or more disconnected (dc-insulated) parts – see, e.g., Fig. 18. (Such structures are more frequently called the transmission lines rather than waveguides, the last term being mostly reserved for the lines with singly-connected cross-sections of the walls.)

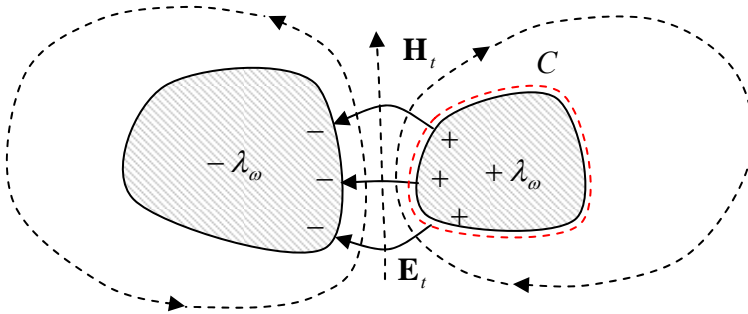


Fig. 7.18. Example of the cross-section of a transmission line that may support the TEM wave propagation.

Now we can readily derive some “global” relations for each conductor, independent on the exact shape of its cross-section. Indeed, consider contour C drawn very close to the conductor’s surface (see, e.g., the red dashed line in Fig. 18). First, we can consider it as a cross-section of a cylindrical Gaussian volume of certain length $dz \ll \lambda \equiv 2\pi/k$. Using the generalized Gauss law (3.29), get

$$\oint_C (\mathbf{E}_t)_n dr = \frac{\lambda_\omega}{\varepsilon}, \quad (7.105)$$

where λ_ω (not to be confused with wavelength λ !) is the linear density of electric charge of the conductor. Second, the same contour C may be used in the generalized Ampère law (5.131) to write

$$\oint_C (\mathbf{H}_t)_\tau dr = I_\omega, \quad (7.106)$$

where I_ω is the total current flowing along the conductor (or rather its complex amplitude). But, as was mentioned above, in the TEM wave the ratio E_t/H_t of the field components participating in these two integrals is constant and equal to $Z = (\mu/\epsilon)^{1/2}$, so that Eqs. (105)-(106) give the following simple relation between the “global” characteristics of the conductor:

$$I_\omega = \frac{\lambda_\omega / \epsilon}{Z} = \frac{\lambda_\omega}{(\epsilon\mu)^{1/2}} = \frac{\omega}{k} \lambda_\omega. \quad (7.107)$$

This relation may be also obtained by a different means; let me describe it, because it has an independent value. Let us consider a small segment $dz \ll \lambda = 2\pi/k$ of the conductor (limited by the red dashed line in Fig. 18) and apply the electric charge conservation law (4.1) to the instant values of the linear charge density and current. The cancellation of dz in both parts yields

$$\frac{\partial \lambda(z, t)}{\partial t} = - \frac{\partial I(z, t)}{\partial z}. \quad (7.108)$$

(If we accept the sinusoidal waveform, $\exp\{i(kz - \omega t)\}$, for both these variables, we immediately recover Eq. (107) for their complex amplitudes, so that the result just expresses the charge continuity law. However, Eq. (108) is valid for any waveform.)

The global equation (108) may be made more specific in the case when the frequency dependence of ϵ and μ is negligible, and the transmission line consists of just two isolated conductors (see, e.g., Fig. 18). In this case, in order to have the wave well localized in the space near the two conductors, we need a sufficiently fast convergence of its electric field at large distances.³⁹ For that, their linear charge densities for each value of z should be equal and opposite, and we can simply relate them to the potential difference V between the conductors:

$$\frac{\lambda(z, t)}{V(z, t)} = C_0, \quad (7.109)$$

where C_0 is the mutual capacitance of the conductors per unit length – that was repeatedly discussed in Chapter 2. Then Eq. (108) takes the form

$$C_0 \frac{\partial V(z, t)}{\partial t} = - \frac{\partial I(z, t)}{\partial z}. \quad (7.110)$$

Next, let us consider the contour shown with the red dashed line in Fig. 19 (which shows a cross-section of the transmission line by a plane containing the wave propagation axis z), and apply to it the Faraday induction law (6.3). Since the electric field is zero inside the conductors (in Fig. 19, on the horizontal parts of the contour), the total e.m.f. equals the difference of voltages V at the end of the segment dz , while the only source of the magnetic flux through the area limited by the contour are the (equal and opposite) currents $\pm I$ in the conductors, we can use Eq. (5.70) to express it. As a result, canceling dz in both parts of the equation, we get

$$L_0 \frac{\partial I(z, t)}{\partial t} = - \frac{\partial V(z, t)}{\partial z}, \quad (7.111)$$

³⁹ The alternative is to have a virtually plane wave, which propagates along the transmission line conductors, and whose fields are just slightly deformed in their vicinity. Such a wave cannot be “guided” by the conductors, and hardly deserves the name of a “wave in the waveguide”.

where L_0 is the mutual inductance of the conductors per unit length. The only difference between L_0 and the dc mutual inductances discussed in Chapter 5 is that at the high frequencies we are analyzing now, L_0 should be calculated neglecting its penetration into the conductors. (In the dc case, we had the same situation for superconductor electrodes, within their crude, ideal-diamagnetic description.)

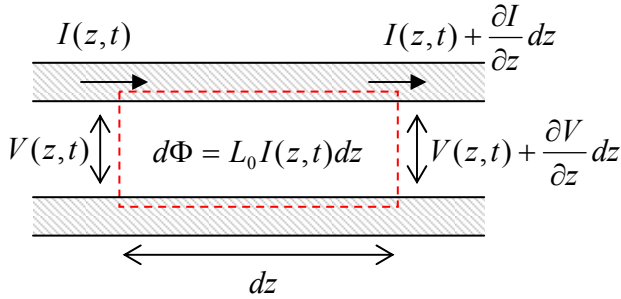


Fig. 7.19. Electric current, magnetic flux, and voltage in a two-conductor transmission line.

The system of Eqs. (110) and (111) is frequently called the *telegrapher's equations*. Combined, they give for any “global” variable f (either V , or I , or λ) a 1D wave equation,

$$\frac{\partial^2 f}{\partial z^2} - L_0 C_0 \frac{\partial^2 f}{\partial t^2} = 0, \quad (7.112)$$

which describes the dispersion-free TEM wave propagation. Again, this equation is only valid within the frequency range where the frequency dependence of both ε and μ is negligible. If it is not so, the global approach may still be used for sinusoidal waves $f = \text{Re}[f_\omega \exp\{i(kz - \omega t)\}]$. Repeating the above arguments, instead of Eqs. (110)-(111) we get algebraic equations

$$\omega C_0 V_\omega = k I_\omega, \quad \omega L_0 I_\omega = k V_\omega, \quad (7.113)$$

in which $L_0 \propto \mu$ and $C_0 \propto \varepsilon$ may now depend on frequency.

Two linear equations (113) are consistent only if

$$L_0 C_0 = \frac{k^2}{\omega^2} \equiv \frac{1}{v^2} \equiv \varepsilon \mu. \quad (7.114)$$

$L_0 C_0$
product
invariance

Besides the fact we have already known (that the TEM wave speed is the same as that of the plane wave), Eq. (114) gives us a result that I confess I have not emphasized enough in Chapter 5: the product $L_0 C_0$ does not depend on the shape or size of line's cross-section (provided that the magnetic field penetration into the conductors is negligible). Hence, if we have calculated the mutual capacitance C_0 of a system of two cylindrical conductors, the result immediately gives us their mutual inductance: $L_0 = \varepsilon \mu / C_0$. This relation stems from the fact that both the electric and magnetic fields may be expressed via the solution of a 2D Laplace equation for system's cross-section.

With Eq. (114) satisfied, any of Eqs. (113) gives the same result for ratio

$$Z_W \equiv \frac{V_\omega}{I_\omega} = \left(\frac{L_0}{C_0} \right)^{1/2}, \quad (7.115)$$

Transmission
line's TEM
Impedance

that is called the *transmission line's impedance*. This parameter has the same dimensionality (in SI units, ohms) as the wave impedance (7),

$$Z \equiv \frac{E_\omega}{H_\omega} = \left(\frac{\mu}{\varepsilon} \right)^{1/2}, \quad (7.116)$$

but these parameters should not be confused, because Z_W depends on cross-section's geometry, while Z does not. In particular, Z_W is the only important parameter of a transmission line for matching with a lumped load circuit (Fig. 20) in the important case when both the cable cross-section's size and the load's linear dimensions are much smaller than the wavelength. (The ability of TEM lines to have such a small cross-section is their another important advantage.) Indeed, in this case we may consider the load in the quasistatic limit and write

$$V_\omega(z_0) = Z_L(\omega) I_\omega(z_0), \quad (7.117)$$

where $Z_L(\omega)$ is the (generally complex) impedance of the load. Taking $V(z,t)$ and $I(z,t)$ in the form similar to Eqs. (61) and (62), and writing two Kirchhoff's laws for point $z = z_0$, we get for the reflection coefficient a result similar to Eq. (68):

$$R = \frac{Z_L(\omega) - Z_W}{Z_L(\omega) + Z_W}. \quad (7.118)$$

This formula shows that for the perfect matching (i.e. the total wave absorption in the load), load's impedance $Z_L(\omega)$ should be real and equal to Z_W - but not necessarily to Z .

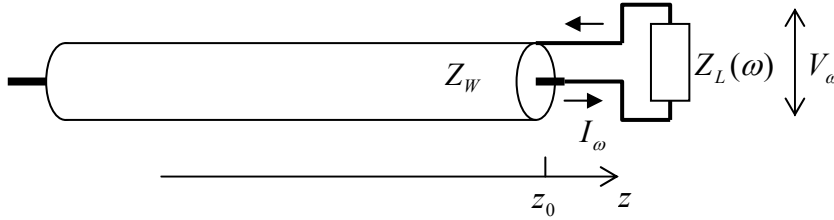


Fig. 7.20. Transmission line impedance matching.

As an example, let us consider one of the simplest (and the most important) transmission lines: the coaxial cable (Fig. 21).⁴⁰

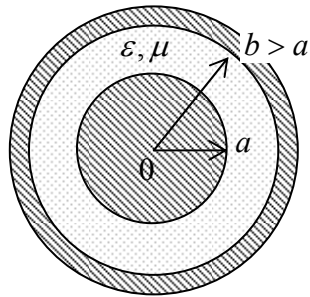


Fig. 7. 21. Cross-section of a coaxial cable with arbitrary (possibly, dispersive) dielectric filling.

For this geometry, we already know expressions for both L_0 and C_0 , though they have to be modified for the dielectric constant and the magnetic field non-penetration into the conductors. After that modification,

⁴⁰ The coaxial cable was first patented by O. Heaviside in 1880.

$$C_0 = \frac{2\pi\epsilon}{\ln(b/a)}, \quad L_0 = \frac{\mu}{2\pi} \ln(b/a). \quad (7.119)$$

Coaxial
cable's
 C_0 and L_0

So, the universal relation (114) is indeed valid! For cable's impedance (115), Eqs. (119) yield

$$Z_W = \left(\frac{\mu}{\epsilon}\right)^{1/2} \frac{\ln(b/a)}{2\pi} = Z \frac{\ln(b/a)}{2\pi} \neq Z. \quad (7.120)$$

For standard TV antenna cables (such as RG-6/U, with $b/a \sim 3$, $\epsilon/\epsilon_0 \approx 2.2$), $Z_W = 75$ ohms, while for most computer component connections, cables with $Z_W = 50$ ohms (such as RG-58/U) are prescribed by electronic engineering standards. Such cables are broadly used for transfer of electromagnetic waves with frequencies (limited mostly by cable attenuation; see Sec. 10 below) up to 1 GHz over distances of a few km, and up to ~ 20 GHz on the tabletop scale (a few meters).

Another important example of TEM transmission lines is the set of two parallel wires. In the form of *twisted pairs*,⁴¹ they allow communications, in particular long-range telephone and DSL Internet connections, at frequencies up to a few hundred kHz, as well as relatively short Ethernet and TV cables at frequencies up to ~ 1 GHz, limited mostly by the mutual interference and parasitic radiation effects.

7.7. H and E waves in metallic waveguides

Let us now return to Eqs. (100) and explore the TE and TM waves - with, respectively, either H_z or E_z different from zero. At the first sight, they may seem more complex. However, equations (101), which determine the distribution of these longitudinal components over the cross-section, are just 2D Helmholtz equations for scalar functions. For simple cross-section geometries may be solved using the methods discussed for the Laplace equation in Chapter 2, in particular the variable separation. After the solution of such an equation has been found, the transverse components of the fields may be calculated by differentiation, using the simple formulas,

$$\mathbf{E}_t = \frac{i}{k_t^2} [k_z \nabla_t E_z - kZ(\mathbf{n}_z \times \nabla_t H_z)], \quad \mathbf{H}_t = \frac{i}{k_t^2} \left[k_z \nabla_t H_z + \frac{k}{Z} (\mathbf{n}_z \times \nabla_t E_z) \right], \quad (7.121)$$

which follow from the two equations in the first line of Eqs. (100).⁴²

In comparison with the electro- and magnetostatics problems, the only conceptually new feature of Eqs. (101), with appropriate boundary conditions, is that they form the so-called *eigenproblems*, with typically many solutions (*eigenfunctions*), each describing a specific wave mode, and corresponding to a specific *eigenvalue* of parameter k_t . The good news here is that these values of k_t are determined by this 2D boundary problem and hence do not depend on k_z . As a result, the dispersion law $\omega(k_z)$ of each mode, that follows from the last form of Eq. (102),

$$\omega = \left(\frac{k_z^2 + k_t^2}{\epsilon\mu} \right)^{1/2} = (v^2 k_z^2 + \omega_c^2)^{1/2}, \quad (7.122)$$

Universal
dispersion
relation

⁴¹ The twisting reduces mutual induction ("crosstalk") between the lines, and parasitic radiation at their bends.

⁴² For that, one of these two linear equations should be first vector-multiplied by \mathbf{n}_z . Note that this approach could not be used to analyze TEM waves, because for them $k_t = 0$, $E_z = 0$, $H_z = 0$, and Eqs. (121) yield uncertainty.

is functionally the same as that of plane waves in a plasma (see Eq. (38), Fig. 6, and their discussion), with the only differences that c is now replaced with $v = 1/(\epsilon\mu)^{1/2}$, the speed of plane (or any TEM) waves in the medium filling the waveguide, and ω_p is replaced with the so-called *cutoff frequency*

$$\omega_c \equiv vk_t, \quad (7.123)$$

specific for each mode. (As Eq. (101) implies, and as we will see from several examples below, k_t has the order of $1/a$, where a is the characteristic dimension of waveguide's cross-section, so that the critical value of the free-space wavelength is of the order of a .) Below the cutoff frequency of each particular mode, it cannot propagate in the waveguide.⁴³ As a result, modes with the *lowest* values of ω_c present special practical interest, because the choice of the signal frequency ω between two lowest values of cutoff frequency guarantees that the waves propagate in the form of only one mode, with the lowest k_t . Such a choice allows to simplify the excitation of the desired mode by wave generators, and to avoid the parasitic transfer of electromagnetic wave energy to undesirable modes by (unavoidable) small inhomogeneities of the system.

The boundary conditions for the Helmholtz equations (101) depend on the propagating wave type. For TM waves (i.e. E modes, with $H_z = 0$ but $E_z \neq 0$), in the macroscopic approximation the boundary condition $E_\tau = 0$ immediately gives

$$E_z|_C = 0, \quad (7.124)$$

where C is the contour limiting the conducting wall's cross-section. For TE waves (the H modes, with $E_z = 0$ but $H_z \neq 0$), the boundary condition is slightly less obvious and may be obtained using, for example, the second equation of system (100), vector-multiplied by \mathbf{n}_z . Indeed, for the component perpendicular to the conductor surface the equation gives

$$ik_z(\mathbf{H}_t)_n - i\frac{k}{Z}(\mathbf{n}_z \times \mathbf{E}_t)_n = \frac{\partial H_z}{\partial n}. \quad (7.125)$$

But the first term in the left-hand part of this equation must be zero on the wall surface, because of the second of Eqs. (103), while according to the first of Eqs. (103), vector \mathbf{E}_t in the second term cannot have a component tangential to the wall. As a result, the vector product in that term cannot have a normal component, so that the term should equal zero as well, and Eq. (125) is reduced to

$$\frac{\partial H_z}{\partial n}|_C = 0. \quad (7.126)$$

Let us see what does this approach give for a simple but practically important example of a metallic-wall waveguide with a rectangular cross-section. In this case it is natural to use the Cartesian coordinates shown in Fig. 22, so that both Eqs. (101) take the simple form

⁴³ An interesting recent twist in the ideas of electromagnetic metamaterials (mentioned in Sec. 5 above) is the so-called ϵ -near-zero materials, designed to have the effective product $\epsilon\mu$ much lower than $\epsilon_0\mu_0$ within certain frequency ranges. Since at these frequencies the speed v (4) becomes much lower than c , the cutoff frequency (123) virtually vanishes. As a result, waves may “tunnel” through very narrow sections of metallic waveguides filled with such materials – see, e.g., M. Silveirinha and N. Engheta, *Phys. Rev. Lett.* **97**, 157403 (2006).

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + k_t^2 \right) f = 0, \quad f = \begin{cases} E_z, & \text{for TM waves,} \\ H_z, & \text{for TE waves.} \end{cases} \quad (7.127)$$

From Chapter 2 we know that the most effective way of solution of such equations in a rectangular region is the variable separation, in which the general solution is represented as a sum of partial solutions of the type

$$f = X(x)Y(y). \quad (7.128)$$

Plugging this expression into Eq. (127), and dividing each term by XY , we get the equation,

$$\frac{1}{X} \frac{d^2 X}{dx^2} + \frac{1}{Y} \frac{d^2 Y}{dy^2} + k_t^2 = 0, \quad (7.129)$$

that should be satisfied for all values of x and y within the waveguide's interior. This is only possible if each term of the sum equals a constant. Taking the X -term and Y -term constants in the form $(-k_x^2)$ and $(-k_y^2)$, respectively, and solving the corresponding ordinary differential equations,⁴⁴ for eigenfunction (128) we get

$$f = (c_x \cos k_x x + s_x \sin k_x x)(c_y \cos k_y y + s_y \sin k_y y), \quad \text{with } k_x^2 + k_y^2 = k_t^2, \quad (7.130)$$

where constants c and s should be found from the boundary conditions. Here the difference between the H modes and E modes pitches in.

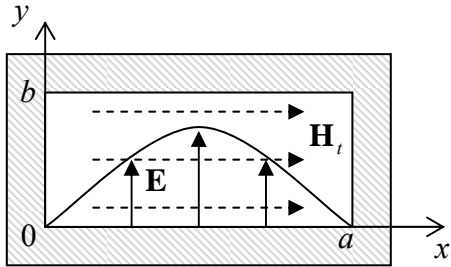


Fig. 7.22. Rectangular waveguide, and the transverse field distribution in the basic mode H_{10} (schematically).

For the former modes (TE waves), Eq. (130) is valid for H_z , and we should use condition (126) on all metallic walls of the waveguide ($x = 0$ and a ; $y = 0$ and b – see Fig. 22). As a result, we get very simple expressions for eigenfunctions and eigenvalues:

$$(H_z)_{nm} = H_l \cos \frac{\pi n x}{a} \cos \frac{\pi m y}{b}, \quad (7.131)$$

$$k_x = \frac{\pi n}{a}, \quad k_y = \frac{\pi m}{b}, \quad (k_t)_{nm} = (k_x^2 + k_y^2)^{1/2} = \pi \left[\left(\frac{n}{a} \right)^2 + \left(\frac{m}{b} \right)^2 \right]^{1/2}, \quad (7.132)$$

⁴⁴ Let me hope that the solution of equations of the type $d^2 X / dx^2 + k_x^2 X = 0$ does not present a problem for the reader, due to his or her prior experience with problems such as standing waves on a guitar string, wavefunctions in a flat 1D quantum well, or (with the replacement $x \rightarrow t$) a classical harmonic oscillator.

where H_l is the longitudinal field amplitude, and n and m are two arbitrary integer numbers, besides that they cannot equal to zero simultaneously. (Otherwise, function $H_z(x,y)$ would be constant, so that, according to Eq. (121), the transverse components of the electric and magnetic field would equal zero. As a result, as the last two lines of Eqs. (100) show, the whole field would be zero for any $k_z \neq 0$.) Assuming, for certainty, that $a \geq b$ (as shown in Fig. 22), we see that the lowest eigenvalue of k_t , and hence the lowest cutoff frequency (123), is achieved for the so-called H_{10} mode with $n = 1$ and $m = 0$, and hence

Basic
mode's
cutoff

$$(k_t)_{10} = \frac{\pi}{a} \quad (7.133)$$

(thus confirming our prior estimate of k_t).

Depending on the a/b ratio, the second lowest k_t and cutoff frequency belong to either the H_{11} mode with $n = 1$ and $m = 1$:

$$(k_t)_{11} = \pi \left(\frac{1}{a^2} + \frac{1}{b^2} \right)^{1/2} = \left[1 + \left(\frac{a}{b} \right)^2 \right]^{1/2} (k_t)_{10}, \quad (7.134)$$

or to the H_{20} mode with $n = 2$ and $m = 0$:

$$(k_t)_{20} = \frac{2\pi}{a} = 2(k_t)_{10}. \quad (7.135)$$

These values become equal at $a/b = \sqrt{3} \approx 1.7$; in practical waveguides, the a/b ratio is not too far from this value. For example, in the standard X-band waveguide WR90 with $a \approx 2.3$ cm ($f_c \equiv \omega_c/2\pi \approx 6.5$ GHz), $b \approx 1.0$ cm.

Now let us have a fast look at alternative TM waves (E modes). For them, we may still should use the general solution (130) with $f = E_z$, but now with boundary condition (124). This gives us eigenfunctions

$$(E_z)_{nm} = E_l \sin \frac{\pi n x}{a} \sin \frac{\pi m y}{b}, \quad (7.136)$$

and the same eigenvalue spectrum (132) as for the H modes. However, now neither n nor m can be equal to zero; otherwise Eq. (136) would give the trivial solution $E_z(x,y) = 0$. Hence the lowest cutoff frequency of TM waves is provided by the so-called E_{11} mode with $n = 1$, $m = 1$, and the eigenvalue is again given by Eq. (134).

Thus the *basic* (or “fundamental”) H_{10} mode is certainly the most important wave in rectangular waveguides; let us have a better look at its field distribution. Plugging the corresponding solution (131) with $n = 1$ and $m = 0$ into the general Eqs. (121), we easily get

$$(H_x)_{10} = -i \frac{k_z a}{\pi} H_l \sin \frac{\pi x}{a}, \quad (H_y)_{10} = 0, \quad (7.137)$$

$$(E_x)_{10} = 0, \quad (E_y)_{10} = i \frac{ka}{\pi} Z H_l \sin \frac{\pi x}{a}. \quad (7.138)$$

This field distribution is (schematically) shown in Fig. 22. Neither of the fields depends on the vertical coordinate – which is very convenient, in particular, for microwave experiments with small samples.

The electric field has only one (vertical) component that vanishes at the side walls and reaches maximum at waveguide's center; its field lines are straight, starting and ending on wall surface charges (whose distribution propagates along the waveguide together with the wave). In contrast, the magnetic field has two nonvanishing components (H_x and H_z), and its field lines are shaped as horizontal loops wrapped around the electric field maxima.

An important question is whether the H_{10} wave may be usefully characterized by a unique impedance introduced similar to Z_W of the TEM modes – see Eq. (115). The answer is *not*, because the main value of Z_W is a convenient description of the impedance matching of the transmission line with a lumped load – see Fig. 20 and Eq. (118). As was discussed above, such simple description is possible (i.e., does not depend on the exact geometry of the connection) only if both dimensions of line's cross-section are much less than λ . But for the H_{10} wave (and more generally, any non-TEM mode) this is impossible – see, e.g., Eq. (129): its lowest frequency corresponds to the TEM wavelength $\lambda_{\max} = 2\pi/(k_t)_{\min} = 2\pi/(k_t)_{10} = 2a$.⁴⁵

Now let us consider metallic waveguides with round cross-section (Fig. 23a). In this single-connected geometry, again, the TEM waves are impossible, while for the analysis of H modes and E modes the polar coordinates $\{\rho, \varphi\}$ are most natural. In these coordinates, the 2D Helmholtz equation (101) takes the form

$$\left[\frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho \frac{\partial}{\partial \rho} \right) + \frac{1}{\rho^2} \frac{\partial^2}{\partial \varphi^2} + k_t^2 \right] f = 0, \quad f = \begin{cases} E_z, & \text{for TM waves,} \\ H_z, & \text{for TE waves.} \end{cases} \quad (7.139)$$

Separating the variables as $f = \mathcal{R}(\rho)\mathcal{A}(\varphi)$, we get

$$\frac{1}{\rho \mathcal{R}} \frac{d}{d\rho} \left(\rho \frac{d\mathcal{R}}{d\rho} \right) + \frac{1}{\rho^2 \mathcal{A}} \frac{d^2 \mathcal{A}}{d\varphi^2} + k_t^2 = 0. \quad (7.140)$$

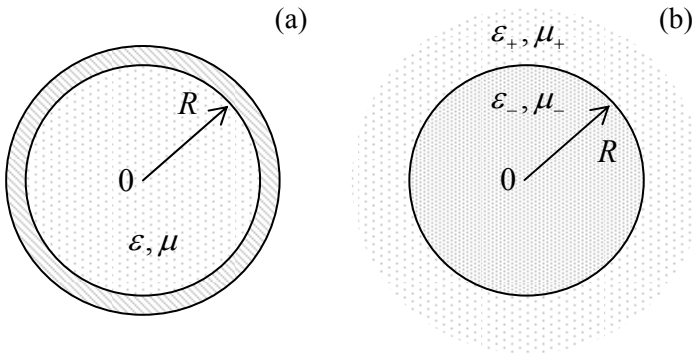


Fig. 7.23. (a) Metallic and (b) dielectric waveguides with circular cross-sections.

But this is exactly the Eq. (2.127) that was studied in the context of electrostatics, just with a replacement of notation: $\gamma \rightarrow k_t$. So we already know that in order to have 2π -periodic functions $\mathcal{A}(\varphi)$, and finite values $\mathcal{R}(0)$ (which are necessary for our current case – see Fig. 23a), the general solution is

⁴⁵ The reader is encouraged to find a simple interpretation of this equality.

given by Eq. (2.136), i.e. the eigenfunctions may be expressed via integer-order Bessel functions of the first kind:⁴⁶

$$f_{nm} = \text{const} \times J_n(k_{nm}\rho)e^{in\varphi}, \quad (7.141)$$

with eigenvalues k_{nm} of the transverse wave number k_t to be determined from appropriate boundary conditions.

As for the rectangular waveguide, let us start from H modes ($f = H_z$). Then the boundary condition on the wall surface ($\rho = R$) is given by Eq. (126), which, for solution (141), takes the form

$$\frac{d}{d\xi} J_n(\xi) = 0, \quad \xi \equiv kR. \quad (7.142)$$

This means that eigenvalues of Eq. (139) are

$$k_t = k_{nm} = \frac{\xi'_{nm}}{R}, \quad (7.143)$$

where ξ'_{nm} is the m^{th} root of function $dJ_n(\xi)/d\xi$. The approximate values of these roots for several lowest n and m may be read out from the plots in Fig. 2.16; their more accurate values are presented in Table 1 below.

Table 7.1. Roots ξ'_{nm} of function $dJ_n(\xi)/d\xi$ for a few values of Bessel function's index n and root's number m .

	$m = 1$	2	3
$n = 0$	3.83171	7.015587	10.1735
1	1.84118	5.33144	8.53632
2	3.05424	6.70613	9.96947
3	4.20119	8.01524	11.34592

It shows, in particular, that the lowest of the roots is $\xi'_{11} \approx 1.84$. Thus, a bit counter-intuitively, the basic mode, providing the lowest cutoff frequency $\omega_c = vk_{nm}$, is H_{11} corresponding to $n = 1$ rather than $n = 0$:⁴⁷

$$H_z = H_1 J_1\left(\xi'_{11} \frac{\rho}{R}\right) e^{i\varphi}, \quad (7.144)$$

with the transverse wave vector $k_t = k_{11} = \xi'_{11}/R \approx 1.84/R$, and hence the cutoff frequency corresponding to the TEM wavelength $\lambda_{\text{max}} = 2\pi/k_{11} \approx 3.41 R$. Thus the ratio of λ_{max} to the waveguide diameter $2R$ is

⁴⁶ In Chapter 2, it was natural to take the angular dependence in the sin-cos form, which is equivalent to adding a similar term with $n \rightarrow -n$ to the right-hand part of Eq. (141). However, since the functions f we are discussing now are already complex, it is easier to do calculations in the exponential form - though it is vital to restore real fields before calculating any of their nonlinear forms, e.g., the wave power.

⁴⁷ The lowest root of Eq. (142) with $n = 0$, i.e. ξ'_{00} , equals 0, and would yield $k = 0$ and hence a constant field H_z , which, according to the first of Eqs. (121), would give vanishing electric field.

about 1.7, i.e. is close to the ratio $\lambda_{\max}/a = 2$ for the rectangular waveguide. The origin of this proximity is clear from Fig. 24, which shows the transverse field distribution in the H_{11} mode. (It may be readily calculated from Eqs. (121) with $E_z = 0$ and H_z given by the real part of Eq. (144).)

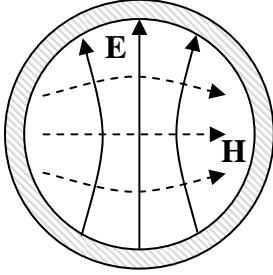


Fig. 7.24. Transverse field components in the basic H_{11} mode of a metallic, circular waveguide (schematically).

One can see that the field structure is actually very similar to that of the basic mode in the rectangular waveguide, shown in Fig. 22, despite the different nomenclature (due to the different type of used coordinates). However, note the arbitrary argument of complex constant H_l in Eq. (144), indicating that in circular waveguides the transverse field polarization is arbitrary. For some practical applications, the degeneracy of these “quasi-linearly-polarized” waves creates problems; they may be avoided by using waves with circular polarization.⁴⁸

As Table 1 shows, the next lowest H mode is H_{21} , for which $k_t = k_{21} = \xi'_{21}/R \approx 3.05/R$, almost twice larger than that of the basic mode, and only then comes the first mode with no angular dependence of the any field, H_{01} , with $k_t = k_{01} = \xi'_{01}/R \approx 3.83/R$.⁴⁹

For the E modes, we may still use Eq. (141) (with $f = E_z$), but with boundary condition (124) at $\rho = R$. This gives the following equation for the problem eigenvalues:

$$J_n(k_{nm}R) = 0, \text{ i.e. } k_{nm} = \frac{\xi_{nm}}{R}, \quad (7.145)$$

where ξ_{nm} is the m -th root of function $J_n(\xi)$ – see Table 2.1. The table shows that the lowest k_t equals to $\xi_{01}/R \approx 2.405/R$. Hence the corresponding mode (E_{01}), with

$$E_z = E_l J_0(\xi_{01} \frac{\rho}{R}), \quad (7.146)$$

has the second lowest cutoff frequency, approximately 30% higher than that of the basic mode H_{11} .

Finally, let us discuss one more topic of general importance – the number N of electromagnetic modes that may propagate in a waveguide within a certain range of relatively large frequencies $\omega \gg \omega_c$. This is easy to calculate for a rectangular waveguide, with its simple expressions (132) for the eigenvalues of $\{k_x, k_y\}$. Indeed, these expressions describe a rectangular mesh on the $[k_x, k_y]$ plane, so

⁴⁸ Actually, Eq. (144) does describe a circularly polarized wave, while the real and imaginary parts of this expression describing two mutually perpendicular quasi-linearly-polarized waves.

⁴⁹ Electric field lines in the H_{01} mode (as well as all higher H_{0m} modes) are directed straight from the axis to the walls, reminding those of TEM waves in the coaxial cable. Due to this property, these modes provide, at $\omega \gg \omega_c$, much lower power losses (see Sec. 10 below) than the fundamental H_{11} mode, and are sometimes used in practice, despite all inconveniences of working in the multimode frequency range.

that each point corresponds to the plane area $\Delta A_k = (\pi/a)(\pi/b)$, and the number of modes in a large k -plane area $A_k \gg \Delta A_k$ is $N = A_k/\Delta A_k = abA_k/\pi^2 = AA_k/\pi^2$, where A is the waveguide's cross-section area.⁵⁰ However, it is frequently more convenient to discuss transverse wave vectors \mathbf{k}_t of arbitrary direction, i.e. with arbitrary sign their components k_x and k_y . Taking into account that the opposite values of each component actually give the same wave, the actual number of different modes of each type (E or H) is a factor of 4 lower than was calculated above. This means that the number of modes of *both* types is

$$N = 2 \frac{A_k A}{(2\pi)^2}. \quad (7.147)$$

It may be convincingly argued that this *mode counting rule* is valid for waveguides with cross-section of any shape, and any boundary conditions on the walls, provided that $N \gg 1$.

7.8. Dielectric waveguides and optical fibers

Now let us discuss electromagnetic wave propagation in *dielectric waveguides*. The simplest, *step-index* waveguide (Figs. 23, 25) consists of an inner *core* and an outer shell (in the optical fiber technology, called *cladding*) with a higher wave propagation speed, i.e. lower index of refraction:

$$v_+ > v_-, \quad \text{i.e. } k_+ < k_-, \quad \varepsilon_+ \mu_+ < \varepsilon_- \mu_-. \quad (7.148)$$

(In most cases the difference is achieved due to that in the dielectric constant, $\varepsilon_- < \varepsilon_+$, while magnetically both materials are almost passive: $\mu_- \approx \mu_+ \approx \mu_0$, and I will assume that in my narrative.) The idea of the waveguide operation may be readily understood in the case when wavelength λ is much smaller than the characteristic size R of core's cross-section. If this “geometric optics” limit, at the distances of the order of λ from the core-to-cladding interface, which determines the wave reflection, we can consider the interface as a plane. As we know from Sec. 5, if angle θ of plane wave incidence on such an interface is larger than the critical value θ_c specified by Eq. (82), the wave is totally reflected. As a result, the waves launched into the fiber core at such “grazing” angles, propagate inside the core, repeatedly reflected from the cladding – see Fig. 25.

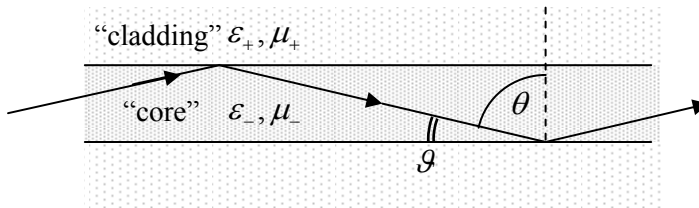


Fig. 7.25. Wave propagation in a thick optical fiber.

The most important type of dielectric waveguides are *optical fibers*.⁵¹ Due to a heroic technological effort, in about three decades starting from the mid-1960s, the attenuation of glass fibers

⁵⁰ This formula ignores the fact that, according to our analysis, some modes (with $n = 0$ and $m = 0$ for H modes, and $n = 0$ or $m = 0$ for E modes, are forbidden. However, for $N \gg 1$, the associated corrections of Eq. (91) are negligible.

⁵¹ For a comprehensive description of this vital technology see, e.g., A. Yariv and P. Yeh, *Photonics*, 6th ed., Oxford U. Press, 2007.

has been decreased from the values of the order of 20 dB/km (typical for the window glass) to the fantastically low values about 0.2 dB/km (meaning a virtually perfect transparency of 10-km-long fiber segments!) – see Fig. 26a. It is remarkable that this ultralow power loss may be combined with an extremely low frequency dispersion, especially for near-infrared waves (Fig. 26b). In conjunction with the development of inexpensive erbium-based quantum amplifiers, this breakthrough has enabled inter-continental (undersea), broadband⁵² optical cables, which are the backbone of all the modern telecommunication infrastructure. The only bad news is that these breakthroughs were achieved for just one kind of materials (silica-based glasses)⁵³ within a very narrow range of their chemical composition. As a result, the dielectric constants $\epsilon_{\pm}/\epsilon_0$ of the cladding and core of practical optical fibers are both close to 2.2 ($n_{\pm} \approx 1.5$) and are very close to each other, so that the relative difference of the refraction indices,

$$\Delta \equiv \frac{n_- - n_+}{n_-} \approx \frac{\epsilon_- - \epsilon_+}{2\epsilon_{\pm}}, \quad (7.149)$$

is typically below 0.5%, thus limiting the fiber bandwidth – see below.

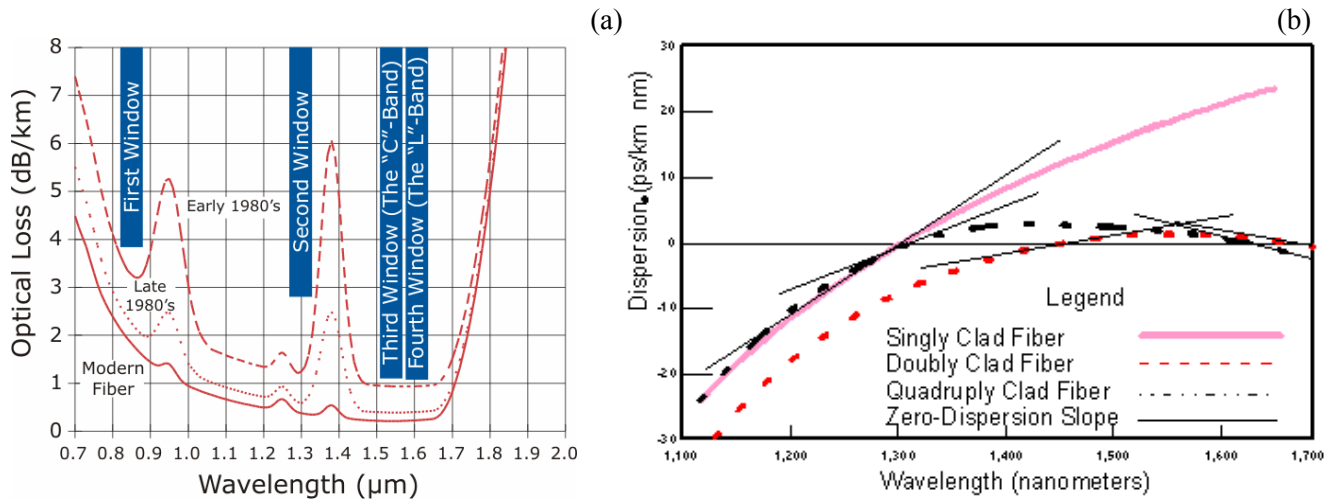


Fig. 7.26. (a) Attenuation and (b) dispersion of representative single-mode optical fibers. (Adapted, respectively, from <http://olson-technology.com> and <http://www.timbercon.com>.)

Practical optical fibers come in two flavors: *multi-mode* and *single-mode* ones. Multi-mode fibers, used for transfer of high optical power (up to as much as ~10 watts), have relatively thick cores, with a diameter $2R$ of the order of 50 μm, much larger than $\lambda \sim 1$ μm. In this case, the “geometric-optics” picture of the wave propagation discussed above is quantitatively correct, and we may use it to calculate the number of quasi-plane-wave modes that may propagate in the fiber. Indeed, for the complementary angle (Fig. 25)

⁵² Each frequency band shown in Fig. 26a, at a typical signal-to-noise ratio $S/N > 10^5$ (50 db), corresponds to the Shannon bandwidth $\Delta f \log_2(S/N)$ exceeding 10^{14} bits per second, five orders of magnitude (!) higher than that of a modern Ethernet cable. And this is only per one fiber; an optical cable may have hundreds of them.

⁵³ The silica-based fibers were suggested in 1966 by C. Kao (the 2009 Nobel Prize in physics), but the very idea of using optical fibers for communications may be traced back to at least the 1963 work by J. Nishizawa.

$$\mathcal{G} \equiv \frac{\pi}{2} - \theta, \quad (7.150)$$

Eq. (82) gives the propagation condition

$$\cos \mathcal{G} > \frac{n_+}{n_-} = 1 - \Delta. \quad (7.151)$$

For the case $\Delta \ll 1$, when the incidence angles $\theta > \theta_c$ of all propagating waves are close to $\pi/2$, and hence the complimentary angles are small, we can keep only two first terms in the Taylor expansion of the left-hand part of Eq. (151) and get

$$\mathcal{G}_{\max}^2 \approx 2\Delta. \quad (7.152)$$

Even for the higher-end value $\Delta = 0.005$, this critical angle is only ~ 0.1 radian, i.e. close to 5° . Due to this smallness, we can approximate the maximum transverse component of the wave vector as

$$(k_t)_{\max} = k(\sin \mathcal{G})_{\max} \approx k\mathcal{G}_{\max} \approx \sqrt{2k\Delta}, \quad (7.153)$$

and use Eq. (147) to calculate number N of propagating modes:

Number
of modes

$$N \approx 2 \frac{(\pi R^2)(\pi k^2 \mathcal{G}_{\max}^2)}{(2\pi)^2} = (kR)^2 \Delta. \quad (7.154)$$

For typical values $k = 0.73 \times 10^7 \text{ m}^{-1}$ (corresponding to the free-space wavelength $\lambda_0 = n\lambda = 2\pi/k \approx 1.3 \text{ } \mu\text{m}$), $R = 25 \text{ } \mu\text{m}$, and $\Delta = 0.005$, this formula gives $N \approx 150$.

The largest problem with using multi-mode fibers for communications is their high *geometric dispersion*, i.e. the difference of the mode propagation speed, which is usually characterized in terms of the signal delay time difference (traditionally measured in picoseconds per kilometer) between the fastest and the slowest mode. Within the geometric optics approximation, the difference of time delays of the fastest mode (with $k_z = k$) and the slowest mode (with $k_z = k \sin \theta_c$) at distance l is

$$\Delta t = \Delta \left(\frac{l}{v_z} \right) = \Delta \left(\frac{k_z l}{\omega} \right) = \frac{l}{\omega} \Delta k_z = \frac{l}{v} (1 - \sin \theta_c) = \frac{l}{v} \left(1 - \frac{n_+}{n_-} \right) = \frac{l}{v} \Delta. \quad (7.155)$$

For the example considered above, the TEM wave speed $v = c/n \approx 2 \times 10^8 \text{ m/s}$, and the geometric dispersion $\Delta t/l$ is close to 25 ps/m , i.e. $25,000 \text{ ps/km}$. (This means, for example, that a 1-ns pulse, being distributed between the modes, would spread to a ~ 25 -ns pulse after passing a just 1-km fiber segment.) Such disastrous dispersion should be compared with *chromatic dispersion* that is due to the frequency dependence of ε_{\pm} , and has the steepness $(dt/d\lambda)/l$ of the order of 10 ps/km-nm (see the solid pink line in Fig. 26b). One can see that through the whole frequency band ($d\lambda \sim 100 \text{ nm}$) the total chromatic dispersion dt/l is of the order of only $1,000 \text{ ps/km}$.

Due to the large geometric dispersion, the multimode fibers are used for signal transfer over only short distances ($\sim 100 \text{ m}$), while long-range communications are based on single-mode fibers, with thin cores (typically with diameters $2R \sim 5 \text{ } \mu\text{m}$, i. e. of the order of $\lambda/\Delta^{1/2}$). For such structures, Eq. (154) yields $N \sim 1$, but in this case the geometric optics approximation is not quantitatively valid, and we should get back to the Maxwell equations. In particular, this analysis should take into an explicit

account the evanescent wave propagating in the cladding, because its penetration depth may be comparable with R .⁵⁴

Since the cross-section of an optical fiber is not uniform and lacks metallic conductors, the Maxwell equations cannot be exactly satisfied with either a TEM, or a TE, or a TM solutions. Instead, the fibers can carry so-called *HE* and *EH* modes, with both fields having longitudinal components simultaneously. In such modes, both E_z and H_z inside the core ($\rho \leq R$) have the form similar to Eq. (141):

$$f_- = f_l J_n(k_t \rho) e^{in\varphi}, \quad \text{with } k_t^2 = k_-^2 - k_z^2 > 0, \quad k_-^2 \equiv \omega^2 \varepsilon_- \mu_-, \quad (7.156)$$

where amplitudes f_l (i.e., E_l and H_l) may be complex to account for the possible angular shift between these components. On the other hand, for the evanescent wave in the cladding, we may rewrite Eq. (102) as

$$(\nabla^2 - \kappa_t^2) f_+ = 0, \quad \text{with } \kappa_t^2 \equiv k_z^2 - k_+^2 > 0, \quad k_+^2 \equiv \omega^2 \varepsilon_+ \mu_+ \quad (7.157)$$

Figure 27 illustrates the relation between k_t , κ_t , k_z , and k_\pm ; note that the following sum,

$$k_t^2 + \kappa_t^2 \equiv \omega^2 (\varepsilon_- - \varepsilon_+) \mu_0, \quad (7.158)$$

Universal
relation
between
 k_t and κ_t

is fixed (at fixed frequency) and, for typical fibers, very small ($\sim 2\Delta k^2 \ll k^2$). By the way, Fig. 27 shows that neither of k_t and κ_t can be larger than $\omega[(\varepsilon_- - \varepsilon_+) \mu_0]^{1/2} = k\Delta^{1/2}$. In particular, this means that the depth $\delta = 1/\kappa_t$ of wave penetration into the cladding is at least $1/k\Delta^{1/2} = \lambda/2\pi\Delta^{1/2} \gg \lambda/2\pi$. This is why the cladding layers in practical optical fibers are made as thick as $\sim 50 \mu\text{m}$, so that only a negligibly small tail of this evanescent wave field reaches their outer surfaces.

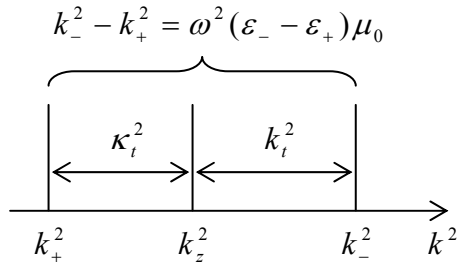


Fig. 7.27. Relation between the transverse exponents k_t and κ_t for waves in optical fibers.

In the polar coordinates, Eq. (157) becomes

$$\left[\frac{1}{\rho} \frac{\partial}{\partial \rho} \left(\rho \frac{\partial}{\partial \rho} \right) + \frac{1}{\rho^2} \frac{\partial^2}{\partial \varphi^2} - \kappa_t^2 \right] f_+ = 0, \quad (7.159)$$

instead of Eq. (139). From Sec. 2.5 we know that the eigenfunctions of Eq. (159) are the products of the angular factor $\exp\{in\varphi\}$ by a linear combination of the modified Bessel functions I_n and K_n , shown in

⁵⁴ I believe that the following calculation is important – both for practice, and as a good example of Maxwell theory application. However, its results will not be used in the following sections/chapters of the course, so that if the reader is not interested in this topic, he or she may safely jump to the beginning Sec. 9.

Fig. 2.20, now of argument $\kappa_t \rho$. In our case, the fields should vanish at $\rho \rightarrow \infty$, so that only the latter functions (of the second kind) can participate:

$$f_+ \propto K_n(\kappa_t \rho) e^{in\varphi}. \quad (7.160)$$

Now we have to reconcile Eqs. (156) and (160), using the boundary conditions at $\rho = R$ for both longitudinal and transverse components of both fields, with the latter fields first calculated from using Eqs. (121). Such a conceptually simple, but a bit bulky calculation (which I am leaving for reader's exercise :-), yields a system of two linear, homogeneous equations for complex amplitudes E_l and H_l , that are compatible if

$$\left(\frac{k_-^2}{k_t} \frac{J_n'}{J_n} + \frac{k_+^2}{\kappa_t} \frac{K_n'}{K_n} \right) \left(\frac{1}{k_t} \frac{J_n'}{J_n} + \frac{1}{\kappa_t} \frac{K_n'}{K_n} \right) = \frac{n^2}{R^2} \left(\frac{k_-^2}{k_t^2} + \frac{k_+^2}{\kappa_t^2} \right) \left(\frac{1}{k_t^2} + \frac{1}{\kappa_t^2} \right), \quad (7.161)$$

where prime means the derivative of each function over its full argument: $k_t \rho$ for J_n , and $\kappa_t \rho$ for K_n .

For any given frequency ω , the system of Eqs. (158) and (161) determines the values of k_t and κ_t , and hence k_z . Actually, for any $n > 0$, this system provides two different solutions: one corresponding to the so-called *HE* wave with larger ratio E_z/H_z , and the *EH* wave, with a smaller value of that ratio. For angular-symmetric modes with $n = 0$ (for whom we might naively expect the lowest cutoff frequency), the equations may be satisfied by fields having just one finite longitudinal component (either E_z or H_z), and the *HE* modes are the usual *E* waves, while the *EH* modes are the *H* waves. For the *H* modes, the characteristic equation is reduced to the requirement that the second parentheses in the left-hand part of Eq. (161) equals to zero. Using the identities $J'_0 = -J_1$ and $K'_0 = -K_1$, this equation may be rewritten as

$$\frac{1}{k_t} \frac{J_1(k_t R)}{J_0(k_t R)} = - \frac{1}{\kappa_t} \frac{K_1(\kappa_t R)}{K_0(\kappa_t R)}. \quad (7.162)$$

Using the simple relation between k_t and κ_t given by Eq. (158), we may plot both parts of Eq. (162) as a function of the same argument, say, $\xi \equiv k_t R$ – see Fig. 28.

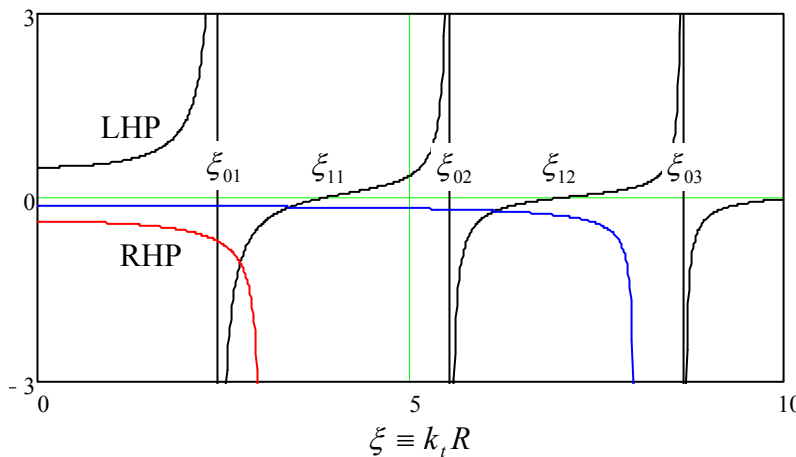


Fig. 7.28. Two sides of the characteristic equation (162), plotted as a function of $k_t R$, for two values of its dimensionless parameter: $V = 8$ (blue line) and $V = 3$ (red line). Note that according to Eq. (158), the argument of functions K_0 and K_1 is just $\kappa_t R = [V^2 - (k_t R)^2]^{1/2} = (V^2 - \xi^2)^{1/2}$.

The right-hand part of Eq. (162) depends not only on ξ but also on the dimensionless parameter V defined as the normalized right-hand part of Eq. (158):

$$V^2 \equiv \omega^2(\varepsilon_- - \varepsilon_+)\mu_0 R^2 \approx 2\Delta k_{\pm}^2 R^2. \quad (7.163)$$

(According to Eq. (155), if $V \gg 1$, it gives the doubled number of the fiber modes – the conclusion confirmed by Fig. 28, taking into account that it describes only the H modes.) Since the ratio K_1/K_0 is positive for all values of their argument (see, e.g., the right panel of Fig. 2.20), the right-hand part of Eq. (162) is always negative, so that the equation may have solutions only in the intervals where the ratio J_l/J_0 is negative, i.e. at

$$\xi_{01} < k_t R < \xi_{11}, \quad \xi_{02} < k_t R < \xi_{12}, \dots, \quad (7.164)$$

where ξ_{nm} is the m -th zero of function $J_n(\xi)$ – see Table 2.1. The right-hand part of the characteristic equation diverges at $k_t R \rightarrow 0$, i.e. at $k_t R \rightarrow V$, so that no solutions are possible if V is below the critical value $V_c = \xi_{01} \approx 2.405$. At this cutoff point, Eq. (163) yields $k_{\pm} \approx \xi_{01}/R(2\Delta)^{1/2}$. Hence, the cutoff frequency for the lowest H mode corresponds to the TEM wavelength

$$\lambda_{\max} = \frac{2\pi R}{\xi_{01}} (2\Delta)^{1/2} \approx 3.7 R \Delta^{1/2}. \quad (7.165)$$

For typical parameters $\Delta = 0.005$ and $R = 2.5 \mu\text{m}$, this result yields $\lambda_{\max} \sim 0.65 \mu\text{m}$, corresponding to the free-space wavelength $\lambda_0 \sim 1 \mu\text{m}$. A similar analysis of the first parentheses in the left-hand part of Eq. (161) shows that at $\Delta \rightarrow 0$, the cutoff frequency for the E modes is similar.

This situation may look exactly like that in metallic waveguides, with no waves possible at frequencies below ω_c , but this is not so. The basic reason for the difference is that in metallic waveguides, the approach to ω_c results in the divergence of the longitudinal wavelength $\lambda_z \equiv 2\pi/k_z$. On the contrary, in dielectric waveguides this approach leaves λ_z finite ($k_z \rightarrow k_+$). Due to this difference, a certain linear superposition of HE and EH modes with $n = 1$ can propagate at frequencies well below the cutoff frequency for $n = 0$, which we have just calculated.⁵⁵ This mode, in the limit $\varepsilon_+ \approx \varepsilon$ (i.e. $\Delta \ll 1$) allows a very interesting and simple description using the *Cartesian* (rather than polar) components of the fields, but still expressed as functions of *polar* coordinates ρ and φ . The reason is that this mode is very close to a linearly polarized TEM wave. (Due to this reason, this mode is referred to as LP_{01} .)

Let us select axis x parallel to the transverse component of the magnetic field vector, so that $E_x|_{\rho=0} = 0$, but $E_y|_{\rho=0} \neq 0$, and $H_x|_{\rho=0} \neq 0$, but $H_y|_{\rho=0} = 0$. The only suitable solutions of the 2D Helmholtz equation (that should be obeyed not only by z -components of the field, but also their x - and y -components) are proportional to $J_0(k_t \rho)$, with zero coefficients for E_x and H_y :

$$E_x = 0, \quad E_y = E_0 J_0(k_t \rho), \quad H_x = H_0 J_0(k_t \rho), \quad H_y = 0, \quad \text{for } \rho \leq R. \quad (7.166)$$

*LP₀₁ mode's
fields
distribution*

Now we can readily calculate the longitudinal components, using the last two equations of Eqs. (100):

$$E_z = \frac{1}{-ik_z} \frac{\partial E_y}{\partial y} = -i \frac{k_t}{k_z} E_0 J_1(k_t \rho) \sin \varphi, \quad H_z = \frac{1}{-ik_z} \frac{\partial H_x}{\partial x} = -i \frac{k_t}{k_z} H_0 J_1(k_t \rho) \cos \varphi, \quad (7.167)$$

where I have used mathematical identities $J'_0 = -J_1$, $\partial \rho / \partial x = x/\rho = \cos \varphi$, and $\partial \rho / \partial y = y/\rho = \sin \varphi$. As a sanity check, we see that the longitudinal component of each field is a (legitimate!) eigenfunction of the

⁵⁵ This fact becomes less surprising if we recall that in the circular metallic waveguide, discussed in Sec. 7, the lowest mode (H_{11} , Fig. 23) also corresponded to $n = 1$ rather than $n = 0$.

type (141) with $n = 1$. Note also that if $k_t \ll k_z$ (this relation is always true if $\Delta \ll 1$ – see Fig. 27), the longitudinal components of the fields are much smaller than their transverse counterparts, so that the wave is indeed very close to the TEM one. Because of that, the ratio of the electric and magnetic field amplitudes is also close to that in the TEM wave: $E_0/H_0 \approx Z_- \approx Z_+$.

Now in order to ensure the continuity of the fields at the core-to-cladding interface ($\rho = R$), we need to have a similar angular dependence of these components at $\rho \geq R$. The longitudinal components of the fields are tangential to the interface and thus should be continuous. Using the solutions similar to Eq. (160) with $n = 1$, we get

$$E_z = -i \frac{k_t}{k_z} \frac{J_1(k_t R)}{K_1(\kappa_t R)} E_0 K_1(\kappa_t \rho) \sin \varphi, \quad H_z = -i \frac{k_t}{k_z} \frac{J_1(k_t R)}{K_1(\kappa_t R)} H_0 K_1(\kappa_t \rho) \cos \varphi, \quad \text{for } \rho \geq R. \quad (7.168)$$

For the transverse components, we should require the continuity of the normal magnetic field μH_n , for our simple field structure equal to just $\mu H_x \cos \varphi$, of the tangential electric field $E_\tau = E_y \sin \varphi$, and of the normal component of $D_n = \varepsilon E_n = \varepsilon E_y \cos \varphi$. Assuming that $\mu = \mu_+ = \mu_0$, and $\varepsilon_+ \approx \varepsilon_-$,⁵⁶ we can satisfy these conditions with the following solutions

$$E_x = 0, \quad E_y = \frac{J_0(k_t \rho)}{K_0(\kappa_t \rho)} E_0 K_0(\kappa_t \rho), \quad H_x = \frac{J_0(k_t \rho)}{K_0(\kappa_t \rho)} H_0 K_0(\kappa_t \rho), \quad H_y = 0, \quad \text{for } \rho \geq R. \quad (7.169)$$

From here, we can calculate components from E_z and H_z , using the same approach as for $\rho \leq R$:

$$\begin{aligned} E_z &= \frac{1}{-ik_z} \frac{\partial E_y}{\partial y} = -i \frac{\kappa_t}{k_z} \frac{J_0(k_t R)}{K_0(\kappa_t R)} E_0 K_1(\kappa_t \rho) \sin \varphi, \\ H_z &= \frac{1}{-ik_z} \frac{\partial H_x}{\partial x} = -i \frac{\kappa_t}{k_z} \frac{J_0(k_t R)}{K_0(\kappa_t R)} H_0 K_1(\kappa_t \rho) \cos \varphi, \quad \text{for } \rho \geq R. \end{aligned} \quad (7.170)$$

We see that this equation provides the same functional dependence of the fields as Eqs. (166), i.e. the internal and external fields are compatible, but their amplitudes coincide only if

$$\boxed{k_t \frac{J_1(k_t R)}{J_0(k_t R)} = \kappa_t \frac{K_1(\kappa_t R)}{K_0(\kappa_t R)}}. \quad (7.171)$$

This characteristic equation (which may be also derived from Eq. (161) with $n = 1$ in the limit $\Delta \rightarrow 0$) looks close to Eq. (162), but functionally is much different from it – see Fig. 29. Indeed, its right-hand part is always positive, and the left-hand part tends to zero at $k_t R \rightarrow 0$. Due to this, Eq. (171) may have a solution for arbitrary small values of parameter V , defined by Eq. (159), i.e. for *arbitrary low frequencies*. This is why this mode is used in practical single-mode fibers: there are no other modes that can propagate at $\omega < \omega_c$, so that the geometric dispersion problem is avoided.

It is easy to use the Bessel function approximations given by the first term of the expansion (2.132) and also Eq. (2.157) to show that in the limit $V \rightarrow 0$ (i.e. $V \ll 1$), $\kappa_t R$ tends to zero much faster

⁵⁶ This is the core assumption of this approximate theory which accounts only for the most important effect of the difference of dielectric constants ε_+ and ε_- : the opposite signs of the differences $(k_+^2 - k_z^2) = k_t^2$ and $(k_-^2 - k_z^2) = -\kappa_t^2$. For more discussion of accuracy of this approximation and some exact results, let me refer the interested reader either to the monograph by A. Snyder and D. Love, *Optical Waveguide Theory*, Chapman and Hill, 1983, or to Chapter 3 and Appendix B in the monograph by Yariv and Yeh, which was cited above.

than $k_t R \approx V$: $\kappa_t R \rightarrow 2\exp\{-1/V\} \ll V$. This means that the scale $\rho_c \equiv 1/\kappa_t$ of the radial distribution of the LP_{01} wave's fields in the cladding becomes very large. In this limit, this mode may be interpreted as a virtually TEM wave propagating in the cladding, just slightly deformed (and guided) by the fiber core. The drawback of this feature is that it requires very thick cladding, in order to avoid energy losses in outer ("buffer" and "jacket") layers that defend the silica components from the elements, but lack their low optical absorption. Due to this reason, the core radius is usually selected so that parameter V is just slightly less than the critical value $V_c = \xi_{01} \approx 2.4$ for higher modes, thus ensuring the single-mode operation and eliminating the geometric dispersion problem.

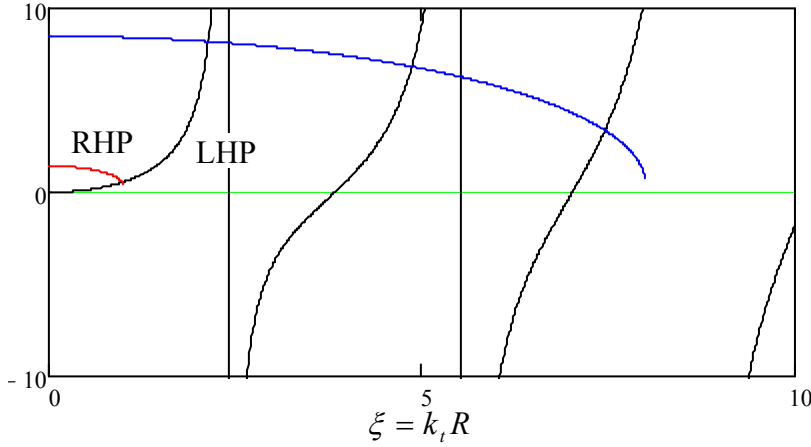


Fig. 7.29. Two sides of the characteristic equation (167) for the LP_{01} mode, plotted as a function of $k_t R$, for two values of the dimensionless parameter: $V = 8$ (blue line) and $V = 1$ (red line).

In order to reduce the field spread into the cladding, the step-index fibers considered above may be replaced with *graded-index* fibers whose the dielectric constant ε_r is gradually and slowly decreased from the center to the periphery. Keeping only the main two terms in the Taylor expansion of the function $\varepsilon(\rho)$ at $\rho = 0$, we may approximate such reduction as

$$\varepsilon(\rho) \approx \varepsilon(0) \left(1 - \frac{\zeta}{2} \rho^2 \right), \quad (7.172)$$

where $\zeta \equiv -[(d^2 \varepsilon / d\rho^2) / \varepsilon]_{\rho=0}$ is a positive constant characterizing the fiber composition gradient.⁵⁷ Moreover, if this constant is sufficiently small ($\zeta \ll k^2$), the field distribution across the fiber's cross-section may be described by the same 2D Helmholtz equation, but with the space-dependent transverse wave vector:⁵⁸

$$[\nabla_t^2 + k_t^2(\rho)]f = 0, \quad \text{where } k_t^2(\rho) \equiv k^2(\rho) - k_z^2 = \omega^2 \varepsilon(\rho) \mu_0 - k_z^2 = k_t^2(0) \left(1 - \frac{\zeta}{2} \rho^2 \right). \quad (7.173)$$

Surprisingly for such axially-symmetric problem, because of its special dependence on the radius, this equation may be most readily solved in Cartesian coordinates. Indeed, rewriting it as

⁵⁷ For an axially-symmetric fiber with a smooth function $\varepsilon(\rho)$, the *first* derivative $d\varepsilon/d\rho$ should vanish at $\rho = 0$.

⁵⁸ Such approach is invalid at arbitrary (large) ζ . Indeed, in the macroscopic Maxwell equations, $\varepsilon(\mathbf{r})$ is under the differentiation sign, and the exact Helmholtz-type equations for fields have additional terms containing $\nabla \varepsilon$.

$$\left[\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + k_t^2(0) \left(1 - \frac{\varsigma}{2} x^2 - \frac{\varsigma}{2} y^2 \right) \right] f = 0, \quad (7.174)$$

and separating variables as $f = X(x)Y(y)$, we get

$$\frac{d^2 X}{dX^2} + \frac{d^2 Y}{dY^2} + k_t^2(0) \left(1 - \frac{\varsigma}{2} x^2 - \frac{\varsigma}{2} y^2 \right) = 0, \quad (7.175)$$

so that functions X and Y obey the same similar differential equation,

$$\frac{d^2 f}{dx^2} + k_x^2 \left(1 - \frac{\varsigma}{2} x^2 \right) f = 0, \quad f = \begin{cases} X, \\ Y. \end{cases} \quad (7.176)$$

with the separation constants satisfying the following relation:

$$k_x^2 + k_y^2 = k_t^2(0) = \omega^2 \varepsilon(0) \mu_0 - k_z^2. \quad (7.177)$$

Equation (176) is well known from the elementary quantum mechanics, because the Schrödinger equation for the perhaps most important quantum system, a 1D harmonic oscillator, may be rewritten in this form. Their eigenvalues are described by a simple formula

$$(k_x)_n = \left(\frac{\varsigma}{2} \right)^{1/2} (2n+1), \quad (k_y)_m = \left(\frac{\varsigma}{2} \right)^{1/2} (2m+1), \quad n, m = 0, 1, 2, \dots \quad (7.178)$$

but eigenfunctions $X_n(x)$ and $Y_m(y)$ have to be expressed via not quite elementary functions - the Hermite polynomials.⁵⁹ For our purposes, however, the lowest eigenfunctions $X_0(x)$ and $Y_0(y)$ are sufficient, because they correspond to the lowest $k_{x,y}$ and hence the lowest cutoff frequency:

$$\omega_c^2 \varepsilon(0) \mu_0 = (k_x)_0^2 + (k_y)_0^2 = \varsigma. \quad (7.179)$$

(Note that at $\varsigma \rightarrow 0$, the cutoff frequency tends to zero, as it should be for a wave in a uniform medium.) The eigenfunctions corresponding to the lowest eigenvalues are simple:

$$f_0(x) = \text{const} \times \exp \left\{ -\frac{\varsigma x^2}{4} \right\}, \quad (7.180)$$

so that the field distribution follows the Gaussian (“bell curve”) function

$$f_0(\rho) = f_0(0) \exp \left\{ -\frac{\varsigma(x^2 + y^2)}{4} \right\} = f_0(0) \exp \left\{ -\frac{\varsigma \rho^2}{4} \right\}. \quad (7.181)$$

This is the so-called *Gaussian beam*, very convenient for some applications. Still, the graded-index fibers have higher attenuation than their step-index counterparts, and are not used as broadly.

Speaking of the Gaussian beams (or more generally, any beams with axially-symmetric profile $f_0(\rho)$), I cannot help noticing the very curious option of forming so-called *helical waves* with complex amplitude $f_0(\rho) \exp\{il\varphi\}$, where l is an integer constant, and φ is the azimuthal angle (so that in our notation $x = \rho \cos \varphi$, $y = \rho \sin \varphi$). Let me leave it for reader's exercise to prove that the electromagnetic

⁵⁹ See, e.g., QM Sec. 2.6.

field of such a wave has an angular momentum vector $\mathbf{L} = L_z \mathbf{n}_z$, with L_z proportional to l .⁶⁰ Quantization of the helical field gives $L_z = l\hbar$ per photon. The case $l = \pm 1$ is possible for infinite-width beams (i.e. plane waves) and means their circular polarization, quantum-mechanically corresponding to spin ± 1 of their photons - see the discussion in the end of Sec. 1. In contrast, the implementation of higher values of $|l|$ requires space-limited beams (with $f_0 \rightarrow 0$ at $\rho \rightarrow \infty$) and may be interpreted as giving the wave an additional “orbital” angular momentum.⁶¹

7.9. Resonators

Resonators are the distributed oscillators, i.e. structures that may sustain standing waves (in electrodynamics, oscillations of the electric and magnetic field at each point) even without a source, until the oscillation amplitude slowly decreases in time due to unavoidable energy losses. If the resonator quality (described by the so-called *Q-factor*, which will be defined and discussed in the next section) is high, this decay takes many oscillation periods. Alternatively, high-*Q* resonators may sustain oscillating fields permanently, if fed with a relatively weak incident wave.

Conceptually the simplest resonator is the *Fabry-Pérot interferometer*⁶² that may be obtained by placing two well-conducting planes parallel to each other.⁶³ Indeed, in Sec. 1 we have seen that if a plane wave is normally incident on such a “perfect mirror”, located at $z = 0$, its reflection, at negligible skin depth, results in a standing wave described by Eq. (61) – that may be rewritten as

$$E(z, t) = \text{Re} \left(2E_\omega e^{-i\omega t + i\pi/2} \right) \sin kz. \quad (7.182)$$

Hence the wave would not change if we had suddenly put the second mirror (isolating the segment of length l from the external wave source) at any position $z = l$ with $\sin kl = 0$, i.e.

$$kl = p\pi, \quad \text{where } p = 1, 2, \dots \quad (7.183)$$

This condition, which also determines the *eigen-* (or *resonance*) *frequency spectrum* of the resonator of fixed length l ,

$$\omega_p = vk_p = \frac{\pi v}{a} p, \quad v = \frac{1}{(\epsilon\mu)^{1/2}}, \quad (7.184)$$

⁶⁰ This task should be easier after reviewing results of field’s momentum analysis in Sec. 9.8, in particular Eqs. (9.235) and (9.237).

⁶¹ Theoretically, the possibility of separating of the angular momentum of an electromagnetic wave to the “spin” and “orbital” parts may be traced back to at least the 1943 work by J. Humblet; however, this issue had not been discussed in literature too much until the spectacular 1992 experiments by L. Allen *et al.* who demonstrated a simple way of generating such helical optical beams. (For reviews of this and later work see, e.g., G. Molina-Terriza *et al.*, *Nature Physics* **3**, 305 (2007) and/or L. Marrucci *et al.*, *J. Opt.* **13**, 064001 (2011), and references therein.) Presently there are efforts to use this approach for so-called “orbital angular moment (OAM) multiplexing” of waves for high-rate information transmission – see, e.g., J. Wang *et al.*, *Nature Photonics* **6**, 488 (2012).

⁶² The device is named after its inventors, M. Fabry and A. Pérot; and is also called the *Fabry-Pérot etalon* (meaning “gauge”), because of its initial usage for the light wavelength measurement.

⁶³ The resonators formed by well conducting (usually, metallic) walls are frequently called the *resonant cavities*.

has a simple physical sense: the resonator length l equals exactly p half-waves of frequency ω_p . Though this is all very simple, please note a considerable change of philosophy from what we have been doing in the previous sections: the main task in resonator analysis is finding its eigenfrequencies ω_p that are now determined by the system geometry rather than by an external wave source.

Before we move to more complex resonators, let us use Eq. (62) to present the magnetic field in the Fabry-Pérot interferometer:

$$H(z, t) = \text{Re} \left(2 \frac{E_\omega}{Z} e^{-i\omega t} \right) \cos kz . \quad (7.185)$$

Expressions (182) and (185) show that in contrast to traveling waves, each field of the standing wave changes simultaneously (proportionately) at all points of the Fabry-Pérot resonator, turning to zero everywhere twice a period. At those instants the electric field energy of the resonator vanishes, but the total energy stays constant, because the magnetic field oscillates (also simultaneously at all points) with the phase shift $\pi/2$. Such behavior is typical for all electromagnetic resonators.

Another, more technical remark is that we can readily get the same results (182)-(185) by solving the Maxwell equations from the scratch. For example, we already know that in the absence of dispersion, losses, and sources, they are reduced to wave equations (3) for any field components. For the Fabry-Pérot resonator's analysis, we can use their 1D form, say, for the transverse component of the electric field:

$$\left(\frac{\partial^2}{\partial z^2} - \frac{1}{v^2} \frac{\partial^2}{\partial t^2} \right) E = 0, \quad (7.186)$$

and solve it as a part of an eigenvalue problem with the corresponding boundary conditions. Indeed, separating time and space variables as $E(z, t) = Z(z) \mathcal{T}(t)$, we get

$$\frac{1}{Z} \frac{d^2 Z}{dz^2} - \frac{1}{v^2} \frac{1}{\mathcal{T}} \frac{d^2 \mathcal{T}}{dt^2} = 0. \quad (7.187)$$

Calling the separation constant k^2 , we get two similar ordinary differential equations,

$$\frac{d^2 Z}{dz^2} + k^2 Z = 0, \quad (7.188)$$

$$\frac{d^2 \mathcal{T}}{dt^2} + k^2 v^2 \mathcal{T} = 0, \quad (7.189)$$

both with sinusoidal solutions, so that their product is a standing wave with a wave vector k and frequency $\omega = kv$, which may be presented by Eq. (182).⁶⁴ Now using the boundary conditions $E(0, t) = E(l, t) = 0$,⁶⁵ we get the eigenvalue spectrum for k_p and hence for $\omega_p = vk_p$, given by Eqs. (183) and (184).

⁶⁴ In this form, the equations are valid even in the presence of dispersion, but with the frequency-dependent wave speed: $v^2 = 1/\epsilon(\omega)\mu(\omega)$.

⁶⁵ This is of course the expression of the first of the general boundary conditions (104). The second if these conditions (for the magnetic field) is satisfied automatically for the transverse waves we are considering.

Lessons from this simple case study may be readily generalized for an arbitrary resonator: there are (at least :-) two methods of finding the eigenfrequency spectrum:

(i) We may look at a traveling wave solution and find where reflecting mirrors may be inserted without affecting the wave's structure. Unfortunately, this method is limited to simple geometries.

(ii) We may solve the general 3D wave equation,

$$\left(\nabla^2 - \frac{1}{v^2} \frac{\partial^2}{\partial t^2} \right) f(\mathbf{r}, t) = 0, \quad (7.190)$$

for field components, as an eigenvalue problem with appropriate boundary conditions. If system parameters (and hence coefficient v) do not change in time, the spatial and temporal variables of Eq. (185) may be *always* separated by taking

$$f(\mathbf{r}, t) = \mathcal{R}(\mathbf{r})\mathcal{T}(t), \quad (7.191)$$

where function $\mathcal{T}(t)$ *always* obeys the same equation (189), having the sinusoidal solution of frequency $\omega = vk$. Plugging this solution back into Eq. (190), for the spatial distribution of the field we get the *3D Helmholtz equation*,

$$(\nabla^2 + k^2) \mathcal{R}(\mathbf{r}) = 0, \quad (7.192)$$

3D
Helmholtz
equation

whose solution (for non-symmetric geometries) may be much more complex.

Let us use these methods to find the eigenfrequency spectrum of a few simple, but practically important resonators. First of all, the first method is completely sufficient for the analysis of any resonator formed as a fragment of a uniform TEM transmission line (e.g., a coaxial cable) between two conducting lids perpendicular to the line direction. Indeed, since in such lines $k_z = k = \omega/v$, and the electric field is perpendicular to the propagation axis, e.g., parallel to the lid surface, the boundary conditions are exactly the same as in the Fabry-Pérot resonator, and we again arrive at the eigenfrequency spectrum (184).

Now let us analyze a slightly more complex system: a rectangular metallic-wall cavity of volume $a \times b \times l$ – see Fig. 30. In order to use the first method, let us consider the resonator as a finite-length ($\Delta z = l$) of the rectangular waveguide stretched along axis z , which was analyzed in detail in Sec. 7. As a reminder, for $a < b$, in the basic H_{10} traveling wave mode, both \mathbf{E} and \mathbf{H} do not depend on y , with vector \mathbf{E} having only y -component. On the contrary, vector \mathbf{H} has both components H_x and H_z , with the phase shift $\pi/2$ between them, with component H_x having the same phase as E_y – see Eqs. (131), (137), and (138). Hence, if a plane, perpendicular to axis z , is placed so that the electric field vanishes on it, H_x also vanishes, so that all the boundary conditions (104) pertinent to a perfect metallic wall are fulfilled simultaneously.

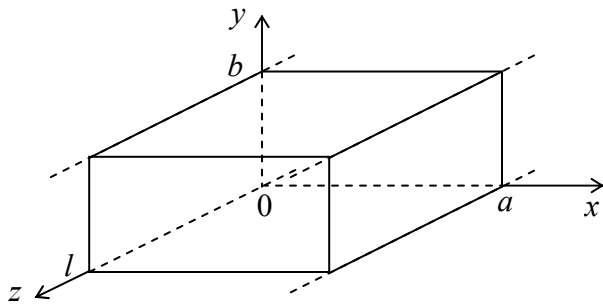


Fig. 7.30. Rectangular metallic resonator as a finite section of a waveguide with the cross-section shown in Fig. 25.

As a result, the H_{10} wave would not be perturbed by two metallic walls separated by an integer number of half-wavelength $\lambda_z/2$ corresponding to the wave number given by Eqs. (102) and (133):

$$k_z = (k^2 - k_t^2)^{1/2} = \left(\frac{\omega^2}{v^2} - \frac{\pi^2}{a^2} \right). \quad (7.193)$$

Using this expression, we see that the smallest of these distances, $l = \lambda_z/2 = \pi/k_z$, gives resonance frequency⁶⁶

$$\omega_{101} = v \left[\left(\frac{\pi}{a} \right)^2 + \left(\frac{\pi}{l} \right)^2 \right]^{1/2}, \quad (7.194)$$

with the indices showing the number of half-waves along each dimension of the system. This is the lowest (fundamental) eigenfrequency of the resonator (if $b < a, l$).

The field distribution in this mode is close to that in the corresponding waveguide mode H_{10} (Fig. 22), with the important difference that phases of the magnetic and electric fields are shifted by phase $\pi/2$ both in space and time, just as in the Fabry-Pérot resonator – see Eqs. (182) and (185). Such time shift allows for a very simple interpretation of the H_{101} mode that is especially adequate for very flat resonators, with $b \ll a, l$. At the instant when the electric field reaches maximum (Fig. 31a), i.e. the magnetic field vanishes in the whole volume, the surface electric charge of the walls (with density $\sigma = E_n/\epsilon$) is largest, being localized mostly in the middle of the broadest (in Fig. 31, horizontal) faces of the resonator. At later times, the walls start to recharge via surface currents whose density J is largest in the side walls, and reaches its maximal value in a quarter period of the oscillation period of frequency ω_{101} – see Fig. 31b. The currents generate the vortex magnetic field, with looped field lines in the plane of the broadest face. The surface currents continue to flow in this direction until (in one more quarter period) the broader walls of the resonator are fully recharged in the polarity opposite to that shown in Fig. 31a. After that, the surface currents start to flow in the direction opposite to that shown in Fig. 31b. This process, that repeats again and again, is conceptually similar to the well-known oscillations in a lumped LC circuit, with the role of (now, distributed) capacitance played mostly by the broadest faces of the resonator, and that of distributed inductance, mostly by its narrower walls.

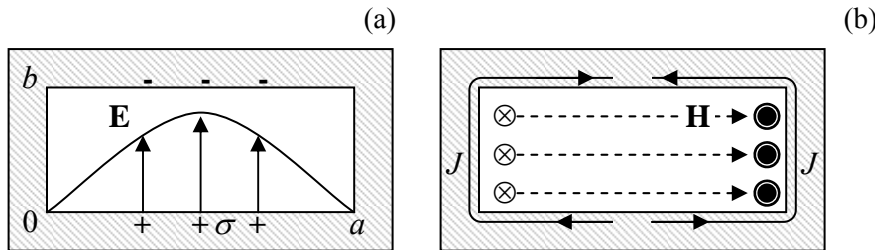


Fig. 7.31. Fields, charges, and currents in the basic H_{101} mode of a rectangular metallic resonator, at two instants separated by $\Delta t = \pi/2\omega_{101}$ – schematically.

In order to generalize result (194) to higher oscillation modes, the second method discussed above is more prudent. Separating variables as $\vec{\mathcal{A}}(\mathbf{r}) = X(x)Y(y)Z(z)$ in the Helmholtz equation (192), we

⁶⁶ In most electrical engineering handbooks, the index corresponding to the shortest side of the resonator is listed last, so that the fundamental mode is nominated as H_{110} and its eigenfrequency as ω_{110} .

see that X , Y , and Z have to be sinusoidal functions of their arguments, with wave vector components satisfying the characteristic equation

$$k_x^2 + k_y^2 + k_z^2 = k^2 \equiv \frac{\omega^2}{v^2}. \quad (7.195)$$

In contrast to the wave propagation problem, now we are dealing with standing waves along all three dimensions, and have to satisfy the boundary conditions on all sets of parallel walls. It is straightforward to check that the macroscopic boundary conditions ($E_\tau = 0$, $H_n = 0$) are fulfilled at the following field component distribution:

$$\begin{aligned} E_x &= E_1 \cos k_x x \sin k_y y \sin k_z z, & H_x &= H_1 \sin k_x x \cos k_y y \cos k_z z, \\ E_y &= E_2 \sin k_x x \cos k_y y \sin k_z z, & H_y &= H_2 \cos k_x x \sin k_y y \cos k_z z, \\ E_z &= E_3 \sin k_x x \sin k_y y \cos k_z z, & H_z &= H_3 \cos k_x x \cos k_y y \sin k_z z, \end{aligned} \quad (7.196)$$

with each of the wave vector components having the equidistant spectrum similar to the one given by Eq. (193):

$$k_x = \frac{\pi n}{a}, \quad k_y = \frac{\pi m}{b}, \quad k_z = \frac{\pi p}{l}, \quad (7.197)$$

so that the full spectrum of eigenfrequencies is given by the following formula,

$$\omega_{nmp} = vk = v \left[\left(\frac{\pi n}{a} \right)^2 + \left(\frac{\pi m}{b} \right)^2 + \left(\frac{\pi p}{l} \right)^2 \right]^{1/2}, \quad (7.198)$$

which is a natural generalization of Eq. (194). Note, however, that of 3 integers m , n , and p at least two have to be different from zero, in order to keep the fields (196) nonvanishing.

Let us use Eq. (199) to evaluate the number of different modes in a relatively small region $d^3k \ll k^3$ (which is still much larger than the reciprocal volume, $1/V = 1/abl$, of the resonator) of the wave vector space. Taking into account that each eigenfrequency (198), with $nml \neq 0$, corresponds to two field modes with different polarizations,⁶⁷ the argumentation absolutely similar to the one used in the end of Sec. 7 for the 2D case yields

$$dN = 2V \frac{d^3k}{(2\pi)^3}. \quad (7.199)$$

Oscillation
mode
density

This property, valid for resonators of arbitrary shape, is broadly used in classical and quantum statistical physics,⁶⁸ in the following form. If some electromagnetic mode property, $f(\mathbf{k})$, is a smooth function of the wave vector, and volume V is large enough, then Eq. (199) may be used to approximate the sum over the modes by an integral:

⁶⁷ This fact becomes evident from plugging Eq. (196) into the Maxwell equation $\nabla \cdot \mathbf{E} = 0$. The resulting equation, $k_x E_1 + k_y E_2 + k_z E_3 = 0$, with the discrete, equidistant spectrum (197) for each wave vector component, may be satisfied by two linearly independent sets of constants $E_{1,2,3}$.

⁶⁸ See, e.g., QM Sec. 1.1 and SM Sec. 2.6.

$$\sum_{\mathbf{k}} f(\mathbf{k}) \approx \int_N f(\mathbf{k}) dN = \int_{\mathbf{k}} f(\mathbf{k}) \frac{dN}{d^3k} d^3k = 2 \frac{V}{(2\pi)^3} \int_{\mathbf{k}} f(\mathbf{k}) d^3k . \quad (7.200)$$

Finally, note that low-loss resonators may be also formed by finite-length sections of not only metallic waveguides with different cross-sections, but also of the dielectric waveguides. Moreover, even the a simple slab of a dielectric material with a μ/ε ratio substantially different from that of its environment (say, the free space) may be used as a high- Q Fabry-Pérot interferometer (Fig. 32), due to an effective wave reflection from its surfaces at normal and especially inclined incidence – see, respectively, Eqs. (68) and Eqs. (91) and (95).

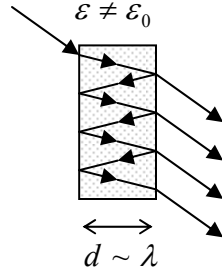


Fig. 7.32. Dielectric Fabry-Pérot interferometer.

Actually, such dielectric Fabry-Pérot interferometer is frequently more convenient for practical purposes than a metallic resonator, due to its natural coupling to environment, that enables a ready way of wave insertion and extraction. The back side of the same medal is that this coupling to environment provides an additional mechanism of power losses, limiting the resonance quality – see the next section.

7.10. Energy loss effects

Inevitable energy losses (“power dissipation”) in passive media lead, in two different situations, to two different effects. In a long transmission line fed by a constant wave source at one end, the losses lead to a gradual *attenuation* of the wave, i.e. to the decrease of its amplitude, and hence power \mathcal{P} , with the distance z along the line. In linear materials, the losses are proportional to the wave amplitude squared, i.e. to the time- average of the power itself, so that the energy balance on a small segment dz takes the form

$$d\mathcal{P} = -\frac{d\mathcal{P}_{\text{loss}}}{dz} dz = -\alpha \mathcal{P} dz . \quad (7.201)$$

Coefficient α , participating in the last form of Eq. (201) and defined by relation

$$\alpha \equiv \frac{d\mathcal{P}_{\text{loss}} / dz}{\mathcal{P}} , \quad (7.202)$$

is called the *attenuation constant*.⁶⁹ Comparing the evident solution of Eq. (201),

⁶⁹ In engineering, attenuation is frequently measured in *decibels per meter* (acronymed as db/m or just dbm):

$$\alpha|_{\text{db/m}} \equiv 10 \log_{10} \frac{\mathcal{P}(z=0)}{\mathcal{P}(z=1 \text{ m})} = 10 \log_{10} e^{\alpha [1/\text{m}]} = \frac{10}{\ln 10} \alpha [\text{m}^{-1}] \approx 4.34 \alpha [\text{m}^{-1}] .$$

$$\mathcal{P}(z) = \mathcal{P}(0)e^{-\alpha z}, \quad (7.203)$$

Wave
attenuation

with Eq. (29), where k is replaced with k_z , we see that α may be expressed as

$$\alpha = 2 \operatorname{Im} k_z, \quad (7.204)$$

where k_z is the component of the wave vector along the transmission line. In the most important limit when the losses are low in the sense $\alpha \ll |k_z| \approx \operatorname{Re} k_z$, its effects on the field distributions along the line's cross-section are negligible, making the calculation of α rather straightforward. In particular, in this limit the contributions to attenuation from two major sources, energy losses in the filling dielectric, and the skin effect in conducting walls, are independent and additive.

The dielectric losses are especially simple to describe. Indeed, a review of our calculations in Secs. 6-8 shows that all of them remain valid if either $\varepsilon(\omega)$, or $\mu(\omega)$, or both, and hence $k(\omega)$, have small imaginary parts:

$$k'' = \omega \operatorname{Im} [\varepsilon^{1/2}(\omega) \mu^{1/2}(\omega)] \ll k'. \quad (7.205)$$

In TEM transmission lines, $k = k_z$, and hence Eq. (205) yields

$$\alpha_{\text{dielectric}} = 2k'' = 2\omega \operatorname{Im} [\varepsilon^{1/2}(\omega) \mu^{1/2}(\omega)]. \quad (7.206)$$

Energy
loss
in filling
dielectric

For dielectric waveguides, in particular optical fibers, these losses are the main attenuation mechanism. As we already know from Sec. 8, in practical optical fibers $\kappa_l R \gg 1$, i.e. most of the field propagates (as the evanescent wave) in the cladding, and the wave mode is very close to TEM. This is why it is sufficient to use Eq. (206) for the cladding material alone.

In waveguides with non-TEM waves, we can readily use the relations between k_z and k derived above to re-calculate k'' into $\operatorname{Im} k_z$. (Note that as such re-calculation, values of k_l stay real, because they are just the eigenvalues of the Helmholtz equation (101), which does not include k .)

In waveguides and transmission lines with metallic conductors, much higher energy losses may come from the skin effect. Let us calculate them, assuming that we know the field distribution in the wave, in particular, the tangential component H of the magnetic field at conductor surface. Then, if the wavelength λ is much larger than δ_s , as it usually is,⁷⁰ we may use the results of the quasistatic approximation derived in Sec. 6.2, in particular Eqs. (6.27)-(6.28) for the relation between the complex amplitudes of the current density in the conductor and the tangential magnetic field

$$j_\omega(x) = k_- H_\omega(x), \quad k_- = -\frac{(1-i)}{\delta_s}, \quad \delta_s^2 = \frac{2}{\mu\omega\sigma}. \quad (7.207)$$

The power loss density (per unit volume) may be now calculated by time averaging of Eq. (4.39):

$$p_{\text{loss}}(x) = \frac{|j_\omega(x)|^2}{2\sigma} = \frac{|k_-|^2 |H_\omega(x)|^2}{2\sigma} = \frac{|H_\omega(x)|^2}{\delta_s^2 \sigma}, \quad (7.208)$$

⁷⁰ As follows from Eq. (78), which may be used for estimates even in cases of arbitrary incidence, this condition is necessary for low attenuation: $\alpha \ll k$ only if $F \ll 1$.

and its integration along the normal to the surface (through all the skin depth), using the exponential law (6.26). This (elementary) integration yields the following power loss per unit area:⁷¹

Energy
loss
in metallic
walls

$$\frac{d\mathcal{P}_{\text{loss}}}{dA} \equiv \int_0^\infty \mathcal{P}_{\text{loss}}(x) dx = |H_\omega(0)|^2 \frac{\mu\omega\delta_s}{4}. \quad (7.209)$$

The total power loss $d\mathcal{P}_{\text{loss}}/dz$ per unit length of a waveguide, i.e. the right-hand part of Eq. (201), now may be calculated by the integration of the ratio $\mathcal{P}_{\text{loss}}/A$ along the contour(s) limiting the cross-section of all conductors of the line. Since our calculation is only valid for low losses, we may ignore their effect on the field distribution, so that the unperturbed distribution may be used both in Eq. (209), i.e. the nominator of Eq. (202), and also for the calculation of the average propagating power, i.e. the denominator of Eq. (202), as the integral of the Poynting vector over the cross-section of the waveguide.

Let us see how this approach works for the TEM mode in one of the simplest TEM transmission lines, the coaxial cable (Fig. 19). As we already know from Sec. 6, in the absence of losses, the distribution of TEM mode fields is the same as in statics, namely:

$$H_z = 0, \quad H_\rho = 0, \quad H_\varphi(\rho) = H_0 \frac{a}{\rho}, \quad (7.210)$$

where H_0 is the field's amplitude on the surface of the inner conductor, and

$$E_z = 0, \quad E_\rho(\rho) = ZH_\varphi(\rho) = ZH_0 \frac{a}{\rho}, \quad E_\varphi = 0, \quad Z \equiv \left(\frac{\mu}{\varepsilon}\right)^{1/2}. \quad (7.211)$$

Now we can, neglecting losses for now, use Eq. (42) to calculate the time-averaged Poynting vector

$$\bar{S} = \frac{Z|H_\varphi(\rho)|^2}{2} = \frac{Z|H_0|^2}{2} \left(\frac{a}{\rho}\right)^2, \quad (7.212)$$

and from it, the total power propagating through the cross-section:

$$\mathcal{P} = \int_A \bar{S} d^2r = \frac{Z|H_0|^2 a^2}{2} 2\pi \int_a^b \frac{\rho d\rho}{\rho^2} = \pi Z |H_0|^2 a^2 \ln \frac{b}{a}. \quad (7.213)$$

For the particular case of the coaxial cable (Fig. 19), the contours limiting the wall cross-sections are circles of radii $\rho = a$ (where the surface field amplitude $H_\omega(0)$ equals, in our notation, H_0), and $\rho = b$ (where, according to Eq. (204), the field is a factor of b/a lower). As a result, for the power loss per unit length, Eq. (209) yields

$$\frac{d\mathcal{P}_{\text{loss}}}{dz} = \left(2\pi a |H_0|^2 + 2\pi b \left| H_0 \frac{a}{b} \right|^2 \right) \frac{\mu_0 \omega \delta_s}{4} = \frac{\pi}{2} a \left(1 + \frac{a}{b} \right) \mu_0 \omega \delta_s |H_0|^2. \quad (7.214)$$

Note that at $a \ll b$, the losses in the inner conductor dominate, despite its smaller surface, because of the higher surface field. Now we may plug Eqs. (213)-(214) into the definition (202) of α , to calculate the part of the attenuation constant associated with the skin effect:

⁷¹ For a normally-incident plane wave, this formula would bring us back to Eq. (78).

$$\alpha_{\text{skin}} \equiv \frac{d\mathcal{P}_{\text{loss}}/dz}{\mathcal{P}} = \frac{1}{2\ln(b/a)} \left(\frac{1}{a} + \frac{1}{b} \right) \frac{\mu\omega\delta_s}{Z} = \frac{k\delta_s}{2\ln(b/a)} \left(\frac{1}{a} + \frac{1}{b} \right). \quad (7.215)$$

We see that the relative (dimensionless) attenuation, α/k , scales approximately as the ratio $\delta_s/\min[a, b]$. This result should be compared with Eq. (78) for the normal incidence of plane waves on a conducting surface.

Let us evaluate α for the standard TV cable RG-6/U (with copper conductors of diameters $2a = 1$ mm, $2b = 4.7$ mm, and $\varepsilon \approx 2.2 \varepsilon_0$, $\mu \approx \mu_0$). According to Eq. (6.27a), for $f = 100$ MHz ($\omega \approx 6.3 \times 10^8$ s⁻¹) the skin depth of pure copper at room temperature (with $\sigma \approx 6.0 \times 10^7$ S/m) is close to 6.5×10^{-6} m, while $k = \omega(\varepsilon\mu)^{1/2} = (\varepsilon/\varepsilon_0)^{1/2}(\omega/c) \approx 3.1$ m⁻¹. As a result, the attenuation is rather low: $\alpha_{\text{skin}} \approx 0.016$ m⁻¹, so that the attenuation length scale $\mathcal{L} \equiv 1/\alpha$ is about 60 m. Hence the attenuation in a cable connecting a roof TV antenna to a TV set in the same house is not a big problem, though using a worse conductor, e.g., steel, would make the losses rather noticeable. (Hence the current worldwide shortage of copper.) However, an attempt to use the same cable in the X-band ($f \sim 10$ GHz) is more problematic. Indeed, though the skin depth $\delta_s \propto \omega^{-1/2}$ decreases with frequency, the wave length drops, i.e. k increases, even faster ($k \propto \omega$), so that the attenuation $\alpha_{\text{skin}} \propto \omega^{1/2}$ becomes close to 0.16 m, and \mathcal{L} to ~ 6 m. This is why at such frequencies, it is more customary to use rectangular waveguides, with their larger internal dimensions $a, b \sim 1/k$, and hence lower attenuation. Let me leave the calculation of this attenuation, using Eq. (209) and the results derived in Sec. 9, for reader's exercise.

The power loss effect on free oscillations *in resonators* is different: there it leads to a gradual decay of oscillation energy \mathcal{E} in time. The useful measure of this decay, called the *Q factor*, may be introduced by writing the temporal analog of Eq. (201):

$$d\mathcal{E} = -\mathcal{P}_{\text{loss}} dt = -\frac{\omega}{Q} \mathcal{E} dt, \quad (7.216)$$

where ω is the eigenfrequency in the loss-free limit, and the dimensional *Q* factor is defined by a relation parallel to Eq. (202):⁷²

$$\frac{\omega}{Q} \equiv \frac{\mathcal{P}_{\text{loss}}}{\mathcal{E}}. \quad (7.217) \quad \text{Q-factor's definition}$$

The solution to Eq. (216),

$$\mathcal{E}(t) = \mathcal{E}(0)e^{-t/\tau}, \quad \text{with } \tau \equiv \frac{Q}{\omega} = \frac{Q/2\pi}{\omega/2\pi} = \frac{QT}{2\pi}, \quad (7.218) \quad \text{Oscillation energy decay}$$

which is an evident temporal analog of Eq. (203), shows the physical meaning of the *Q* factor: the characteristic time τ of the oscillation energy decay is $(Q/2\pi)$ times longer than the oscillation period $T = 2\pi/\omega$. (Another interpretation of *Q* comes from the relation⁷³

$$Q = \frac{\omega}{\Delta\omega}, \quad (7.219) \quad \text{FWHM bandwidth}$$

⁷² As losses grow, the oscillation waveform deviates from sinusoidal one, and the very notion of “oscillation frequency” becomes vague. As a result, parameter *Q* is well defined only if it is much higher than 1.

⁷³ See, e.g., CM Sec. 4.1.

where $\Delta\omega$ is the so-called *FWHM*⁷⁴ bandwidth of the resonance, namely the difference between the two values of the external signal frequency, one above and one below ω , at which the energy of forced oscillations induced in the resonator by an input signal is twice lower than its resonant value.)

In the important particular case of resonators formed by insertion of metallic walls into a TEM transmission line of small cross-section (with the linear size scale a much less than the wavelength λ), there is no need to calculate the Q factor directly if the line attenuation coefficient α is already known. In fact, as was discussed in Sec. 9 above, the standing waves in such a resonator, of the length given by Eq. (183): $l = p(\lambda/2)$ with $p = 1, 2, \dots$, may be understood as an overlap of two TEM waves running in opposite directions, or in other words, a traveling wave and its reflection from one of the ends, the whole roundtrip taking time $\Delta t = 2l/v = p\lambda/v = 2\pi p/\omega = pT$. According to Eq. (201), at this distance the wave's power should drop by $\exp\{-2\alpha l\} = \exp\{-p\alpha\lambda\}$. On the other hand, the same decay may be viewed as happening in time, and according to Eq. (216), result in the drop by $\exp\{-\Delta t/\tau\} = \exp\{-(pT)/(Q/\omega)\} = \exp\{-2\pi p/Q\}$. Comparing these two exponents, we get

Q vs. α

$$Q = \frac{2\pi}{\alpha\lambda} = \frac{k}{\alpha}. \quad (7.220)$$

This simple relation neglects the losses at wave reflection from the walls limiting the resonator length. Such approximation is indeed legitimate at $a \ll \lambda$; if this relation is violated, or if we are dealing with more complex resonator modes (such as those based on the reflection of E or H waves), the Q factor may be smaller than that given by Eq. (220), and needs to be calculated directly. A substantial relief for such a direct calculation is that, just at the calculation of small attenuation in waveguides, in the low-loss limit ($Q \gg 1$), both the nominator and denominator of the right-hand part of Eq. (217) may be calculated neglecting the effects of the power loss on the field distribution in the resonator. I am leaving such a calculation, for the simplest (rectangular and circular) resonators, for reader's exercise.

To conclude this chapter, the last remark: in some resonators (including certain dielectric resonators and metallic resonators with holes in their walls), additional losses due to wave radiation into the environment are also possible. In some simple cases (say, the Fabry-Pérot interferometer shown in Fig. 32) the calculation of these *radiative losses* is straightforward, but sometimes it requires more elaborated approaches, which will be discussed in the next chapter.

7.11. Exercise problems

7.1.* Find the temporal Green's function of a medium whose complex dielectric constant obeys Eq. (32), using:

- (i) the Fourier transform, and
- (ii) the direct solution of Eq. (30), which describes the corresponding model of the medium.

Hint: For the Fourier transform, you may like to use the Cauchy integral.⁷⁵

7.2. The electric polarization of a material responds in the following way to an electric field step:⁷⁶

⁷⁴ This is the acronym for "Full Width at Half-Maximum".

⁷⁵ See, e.g., MA Eq. (15.2).

$$P(t) = \varepsilon_1 E_0 \left(1 - e^{-t/\tau}\right), \quad \text{if } E(t) = E_0 \times \begin{cases} 0, & \text{for } t < 0, \\ 1, & \text{for } 0 < t, \end{cases}$$

where τ is a positive constant.

7.3. Calculate the complex dielectric constant $\varepsilon(\omega)$ for a material whose dielectric-response Green's function, defined by Eq. (23), is

$$G(\theta) = G_0 \left(1 - e^{-\theta/\tau}\right),$$

with some positive constants G_0 and τ . What is the difference between this dielectric response and the apparently similar one considered in the previous problem?

7.4. Use the Lorentz oscillator model of an atom, given by Eq. (30), to calculate the average potential energy of the atom in a uniform, sinusoidal ac electric field, and use the result to calculate the potential profile created for the atom by a standing electromagnetic wave with the electric field amplitude $E_\omega(\mathbf{r})$. Discuss the conditions of validity of your result.

7.5. The solution of the previous problem shows that a standing plane wave exerts a time-averaged force on a non-relativistic charged particle. Reveal the physics of this force by writing and solving the equations of motion of a free particle in:

- (i) a linearly-polarized, monochromatic, plane traveling wave, and
- (ii) a similar but standing wave.

Discuss the conditions of validity of your result.

7.6. Calculate, sketch and discuss the dispersion relation for electromagnetic waves propagating in a Lorentz oscillator medium described by Eq. (32), for the case of negligible damping.

7.7. As was briefly discussed in Sec. 2,⁷⁷ a wave pulse of a finite but relatively large spatial extension $\Delta r \gg \lambda \equiv 2\pi/k$ may be represented with a *wave packet* – a sum of sinusoidal waves with wave vectors \mathbf{k} within a relatively narrow interval. Consider an electromagnetic plane wave packet of this type, with the electric field distribution

$$\mathbf{E}(\mathbf{r}, t) = \text{Re} \int_{-\infty}^{+\infty} \mathbf{E}_k e^{i(kz - \omega_k t)} dk, \quad \text{with } \omega_k [\varepsilon(\omega_k) \mu(\omega_k)]^{1/2} \equiv |k|,$$

propagating along axis z in an isotropic, linear, and loss-free (but not necessarily dispersion-free) medium. Express the full energy of the packet (per unit area of wave's front) via complex amplitudes \mathbf{E}_k , and discuss its dependence of time.

7.8.* Analyze the effect of a constant, uniform magnetic field \mathbf{B}_0 , parallel to the direction \mathbf{n} of electromagnetic wave propagation, on the wave dispersion in plasma, within the same simple model that

⁷⁶ This function $E(t)$ is of course proportional to the well-known step function θ - see, e.g., MA Eq. (14.3). I am not using this notion just to avoid a possible confusion between two different uses of the Greek letter θ .

⁷⁷ And in more detail in CM Sec. 5.3, and especially in QM Sec. 2.1.

was used in the lecture notes for derivation of Eq. (7.38). (Limit your analysis to relatively weak waves, whose magnetic field is negligible in comparison with \mathbf{B}_0 .)

Hint: You may like to represent the incident wave as a linear superposition of two circularly polarized waves, with the left- and right-hand polarization.

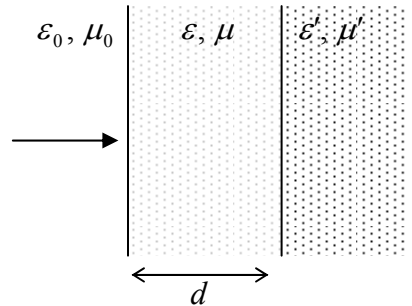
7.9. A monochromatic, plane electromagnetic wave is normally incident from free space on a uniform slab of a material with electric permittivity ε and magnetic permeability μ , with the slab thickness d comparable with the wavelength.

(i) Calculate the power transmission coefficient \mathcal{T} , i.e. the fraction of the incident power, that is transmitted through the slab.

(ii) Assuming that ε and μ are frequency-independent and positive, analyze in detail the frequency dependence of \mathcal{T} . In particular, how does function $\mathcal{T}(\omega)$ depend on the film thickness d and the wave impedance $Z = (\mu/\varepsilon)^{1/2}$ of its material?

7.10. A monochromatic, plane electromagnetic wave with free-space wave number k_0 is normally incident on a plane conducting film of thickness $d \sim \delta_s \ll 1/k_0$. Calculate the power transmission coefficient of the system, i.e. the fraction of incident wave's power propagating beyond the film. Analyze the result in the limits of small and large ratios d/δ_s .

7.11. A plane wave of frequency ω is normally incident, from free space, on a plane surface of a material with real values of the electric permittivity ε' and magnetic permeability μ' . To minimize wave reflection from the surface, you may cover it with a layer, of thickness d , of another transparent material – see Fig. on the right. Calculate the optimal values of ε , μ , and d .

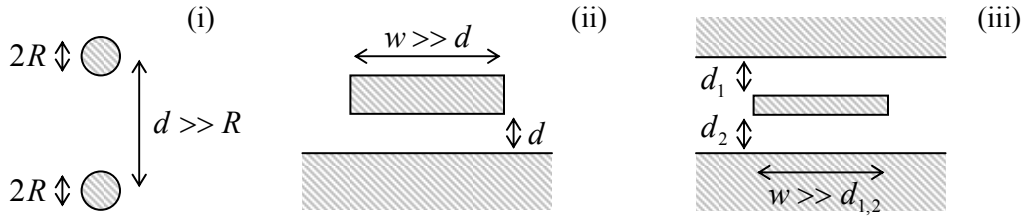


7.12. A monochromatic, plane wave is incident from inside a medium with $\varepsilon\mu > \varepsilon_0\mu_0$ on its plane surface, at the angle of incidence θ larger than the critical angle $\theta_c = \arcsin(\varepsilon_0\mu_0/\varepsilon\mu)^{1/2}$. Calculate the depth δ of the evanescent wave penetration into the free space and analyze its dependence on θ . Does the result depend on the wave polarization?

7.13. Analyze the possibility of propagation of surface electromagnetic waves along a plane boundary between plasma and free space. In particular, calculate and analyze the dispersion relation of the waves.

Hint: Assume that the magnetic field of the wave is parallel to the boundary and perpendicular to the wave propagation direction. (After solving the problem, justify this mode choice.)

7.14. Calculate the characteristic impedance Z_W of the long, straight TEM transmission lines formed by metallic electrodes with cross-sections shown in Fig. below:



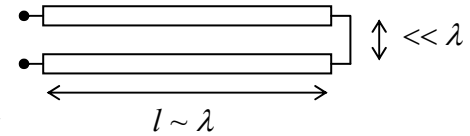
- (i) two round, parallel wires, separated by distance $d \gg R$,
(ii) *microstrip line* of width $w \gg d$,
(iii) *stripline* with $w \gg d_1 \sim d_2$,

in all cases using the macroscopic boundary conditions on metallic surfaces. Assume that the conductors are embedded into a linear dielectric with constant ϵ and μ .

7.15. Modify results of Problem 10(ii) for a superconductor microstrip line, taking into account the magnetic field penetration into both the strip and the ground plane.

7.16.* What lumped ac circuit would be equivalent to the system shown in Fig. 20, with incident wave's power \mathcal{P}_i ? Assume that the wave reflected from the load circuit does not return to it.

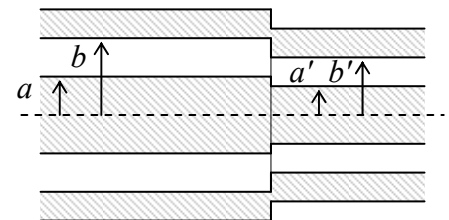
7.17. Find the lumped ac circuit equivalent to a loss-free TEM transmission line of length $l \sim \lambda$, with a small cross-section area $A \ll \lambda^2$, as “seen” (measured) from one end, if the line's conductors are galvanically connected (“shortened”) at the other end – see Fig. on the right. Discuss result's dependence on the signal frequency.



7.18. Represent the fundamental H_{10} wave in a rectangular waveguide (Fig. 22) with a sum of two plane waves, and discuss the physics behind such a representation.

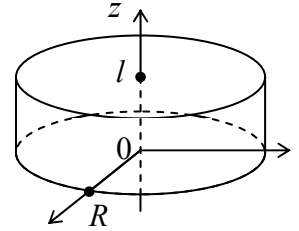
7.19.* For a metallic coaxial cable with the circular cross-section (Fig. 21), find the lowest non-TEM mode and calculate its cutoff frequency.

7.20. Two coaxial cable sections are connected coaxially - see Fig. on the right, which shows system's cut along its symmetry axis. Relations (118) and (120) seem to imply that if the ratios b/a of these sections are equal, their impedance matching is perfect, i.e. a TEM wave incident from one side on the connection would pass it without any reflection at all: $R = 0$. Is this statement correct?



7.21.* Use the recipe outlined in Sec. 8 to prove the characteristic equation (161) for the HE and EH modes in a round, step-index optical fiber.

7.22. Find the lowest eigenfrequencies, and corresponding oscillation modes, of a round cylindrical resonator (see Fig. on the right) with perfectly conducting walls.



7.23. A plane, monochromatic wave propagates through a medium whose Ohmic conductance σ dominates the power losses, while the electric and magnetic polarization effects are negligible. Calculate the wave attenuation coefficient and relate the result with some calculation carried out in Chapter 6.

7.24. Generalize the telegrapher's equations (110)-(111) by taking into account small energy losses in:

- (i) transmission line's conductors, and
- (ii) the media separating the conductors,

using their simplest (Ohmic) models. Formulate the conditions of validity of the resulting equations.

7.25. Calculate the skin-effect contribution to the attenuation coefficient α , defined by Eq. (202), for the basic (H_{10}) mode propagating in a waveguide with the rectangular cross-section – see Fig. 22. Use the results to evaluate α and L for a 10 GHz wave in the standard X-band waveguide WR-90 (with copper walls, $a = 23$ mm, $b = 10$ mm, and no dielectric filling), at room temperature. Compare the estimate with that, made in Sec. 10, for a standard coaxial cable, for the same frequency.

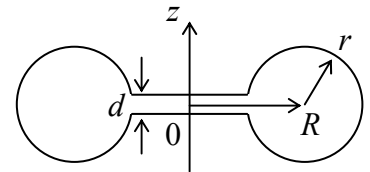
7.26.* Calculate the skin-effect contribution to the attenuation coefficient α of

- (i) the basic (H_{11}) mode, and
- (ii) the H_{01} mode

in a metallic waveguide with the circular cross-section (Fig. 23a), and analyze the low-frequency ($\omega \rightarrow \omega_c$) and high-frequency ($\omega \gg \omega_c$) behaviors of α for each of these modes.

7.27. For a rectangular metallic-wall resonator with dimensions $a \times b \times l$ ($b \leq a, l$), calculate the Q -factor in the fundamental (lowest) oscillation mode, due to the skin-effect losses in the walls. Evaluate the factor (and the lowest eigenfrequency) for a $23 \times 23 \times 10$ mm³ resonator with copper walls, at room temperature.

7.28.* Calculate the lowest eigenfrequency and Q factor (due to the skin-effect losses) of the toroidal (axially-symmetric) resonator with metallic walls and interior's cross-section shown in Fig. on the right, within the limit $d \ll r, R$.



7.29. Express the contribution to the damping coefficient (the reciprocal Q -factor) of a resonator, due to small energy losses in the dielectric that fills it, via dielectric's complex functions $\epsilon(\omega)$ and $\mu(\omega)$ of the material.

7.30. For the dielectric Fabry-Pérot resonator (Fig. 32) with the normal wave incidence, find the Q -factor due to radiation losses in the limit of strong impedance mismatch ($Z \gg Z_0$), using two methods:

- (i) from the energy balance, using Eq. (217), and
- (ii) from the frequency dependence of the power transmission coefficient, using Eq. (219).

Compare the results.

Chapter 8. Radiation, Scattering, Interference, and Diffraction

This chapter continues the discussion of the electromagnetic wave propagation, now focusing on the results of wave incidence on a passive object. Depending on the object's shape, the result of this interaction is called either scattering, or diffraction, or interference. However, as we will see below, the boundary between these effects is blurry, and their mathematical description may be conveniently based on a single key calculation - the electric dipole radiation of a spherical wave by a small source. Naturally, I will start the chapter from this calculation, deriving it from an even more general result – the “retarded potentials” solution of the Maxwell equations.

8.1. Retarded potentials

Let us start from the general solution of the Maxwell equations in a dispersion-free, linear, uniform, isotropic medium, characterized by frequency-independent, real ε and μ - for example, free space.¹ The easiest way to perform this calculation is to use the scalar (ϕ) and vector (\mathbf{A}) potentials of electromagnetic field, that are defined via the electric and magnetic fields by Eqs. (6.106):

$$\mathbf{E} = -\nabla\phi - \frac{\partial\mathbf{A}}{\partial t}, \quad \mathbf{B} = \nabla \times \mathbf{A}. \quad (8.1)$$

As was discussed in Chapter 6, imposing upon the potentials the Lorenz gauge condition (6.108),

$$\nabla \cdot \mathbf{A} + \frac{1}{v^2} \frac{\partial\phi}{\partial t} = 0, \quad v^2 \equiv \frac{1}{\varepsilon\mu}, \quad (8.2)$$

(which does not affect fields \mathbf{E} and \mathbf{B}) the macroscopic Maxwell equations for the fields may be recast into a pair of very similar, simple equations (6.109) for the potentials:

$$\nabla^2\phi - \frac{1}{v^2} \frac{\partial^2\phi}{\partial t^2} = -\frac{\rho}{\varepsilon}, \quad (8.3a)$$

$$\nabla^2\mathbf{A} - \frac{1}{v^2} \frac{\partial^2\mathbf{A}}{\partial t^2} = -\mu\mathbf{j}. \quad (8.3b)$$

Let us calculate the fields induced by the stand-alone electric charge and current densities $\rho(\mathbf{r}, t)$ and $\mathbf{j}(\mathbf{r}, t)$, thinking of them as known functions.² The idea how this may be done may be borrowed from electro- and magnetostatics. Indeed, for the stationary case ($\partial/\partial t = 0$), the solutions of Eqs. (8.3) are given, by the evident generalization of, respectively, Eq. (1.38) and by Eq. (5.28) to the uniform, linear medium:

$$\phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon} \int \rho(\mathbf{r}') \frac{d^3r'}{|\mathbf{r} - \mathbf{r}'|}, \quad (8.4a)$$

¹ When necessary (e.g., at the discussion of the Cherenkov radiation in Sec. 10.4), it will be not too hard to generalize these results to dispersive media.

² Such *thinking* would not prevent the results from being valid for the case when $\rho(\mathbf{r}, t)$ and $\mathbf{j}(\mathbf{r}, t)$ should be calculated self-consistently.

$$\mathbf{A}(\mathbf{r}) \equiv \frac{\mu}{4\pi} \int \mathbf{j}(\mathbf{r}') \frac{d^3 r'}{|\mathbf{r} - \mathbf{r}'|}. \quad (8.4b)$$

As we know, these expressions may be derived by, first, calculating the potential of a point source, and then using the linear superposition principle for a system of such sources.

Let us do the same for the time-dependent case, starting from the field induced by a time-dependent point charge at origin:³

$$\rho(\mathbf{r}, t) = q(t)\delta(\mathbf{r}), \quad (8.5)$$

In this case Eq. (3a) is homogeneous everywhere but the origin:

$$\nabla^2 \phi - \frac{1}{v^2} \frac{\partial^2 \phi}{\partial t^2} = 0, \quad \text{at } r \neq 0. \quad (8.6)$$

Due to the spherical symmetry of the problem, it is natural to look for a spherically-symmetric solution to this equation.⁴ Thus, we may simplify the Laplace operator⁵ correspondingly, and reduce Eq. (6) to

$$\left[\frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) - \frac{1}{v^2} \frac{\partial^2}{\partial t^2} \right] \phi = 0, \quad \text{at } r \neq 0. \quad (8.7)$$

If we now introduce a new variable $\chi \equiv r\phi$, Eq. (7) is reduced to the 1D wave equation

$$\left(\frac{\partial^2}{\partial r^2} - \frac{1}{v^2} \frac{\partial^2}{\partial t^2} \right) \chi = 0, \quad \text{at } r \neq 0. \quad (8.8)$$

From the discussion in Chapter 7,⁶ we know that its general solution may be presented as

$$\chi(r, t) = \chi_{\text{out}} \left(t - \frac{r}{v} \right) + \chi_{\text{in}} \left(t + \frac{r}{v} \right), \quad (8.9)$$

where χ_{in} and χ_{out} are (so far) arbitrary functions of one variable. The physical sense of $\phi_{\text{out}} = \chi_{\text{out}}/r$ is a spherical wave propagating from our source (at $r = 0$) to outer space, i.e. exactly the solution we are looking for. On the other hand, $\phi_{\text{in}} = \chi_{\text{in}}/r$ describes a spherical wave that could be created by some distant spherically-symmetric source, that converges on our charge located at the origin – evidently not the effect we want to consider here. Discarding this term, and returning to $\phi = \chi/r$, we can write the solution (7) as

$$\phi(r, t) = \frac{1}{r} \chi_{\text{out}} \left(t - \frac{r}{v} \right). \quad (8.10)$$

³ Admittedly, this expression does *not* satisfy the continuity equation (4.5), but we will correct this deficiency imminently, at the linear superposition stage – see Eq. (17) below.

⁴ Let me emphasize that this is *not* the general solution to Eq. (6). For example, it does not describe the fields created by other sources, that pass by the considered charge $q(t)$. However, such fields are irrelevant for our current task: to calculate the field created by the charge $q(t)$ *itself*.

⁵ See, e.g., MA Eq. (10.9).

⁶ See also CM Sec. 5.3.

In order to find function χ_{out} , let us consider distances r so small that the time derivative in Eq. (3a), with the right-hand part (5),

$$\nabla^2 \phi - \frac{1}{v^2} \frac{\partial^2 \phi}{\partial t^2} = -\frac{q(t)}{\epsilon} \delta(\mathbf{r}), \quad (8.11)$$

is much smaller than the spatial derivative (that diverges at $r \rightarrow 0$). Then Eq. (11) is reduced to the electrostatic equation whose solution (4a), for source (5), is

$$\phi(r \rightarrow 0, t) = \frac{q(t)}{4\pi\epsilon r}. \quad (8.12)$$

Now requiring the two solutions, (10) and (12), to coincide at $r \ll vt$, we get $\chi_{\text{out}}(t) = q(t)/4\pi\epsilon r$, so that Eq. (10) becomes

$$\phi(r, t) = \frac{1}{4\pi\epsilon r} q\left(t - \frac{r}{v}\right). \quad (8.13)$$

Just as had been done in statics, this result may be readily generalized for the arbitrary position \mathbf{r}' of the point charge:

$$\rho(\mathbf{r}, t) = q(t) \delta(\mathbf{r} - \mathbf{r}') \equiv q(t) \delta(\mathbf{R}), \quad (8.14)$$

where R is the distance between the field observation point \mathbf{r} and the source position point \mathbf{r}' , i.e. the length of the vector,

$$\mathbf{R} \equiv \mathbf{r} - \mathbf{r}', \quad (8.15)$$

connecting these points - see Fig. 1.

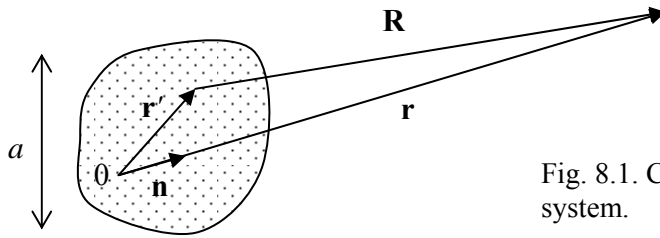


Fig. 8.1. Calculating retarded potentials of a localized system.

Obviously, Eq. (13) becomes

$$\phi(\mathbf{r}, t) = \frac{1}{4\pi\epsilon R} q\left(t - \frac{R}{v}\right). \quad (8.16)$$

Now we can use the linear superposition principle to write, for the arbitrary charge distribution $\rho(\mathbf{r}, t)$,

Retarded
scalar
potential

$$\phi(\mathbf{r}, t) = \frac{1}{4\pi\epsilon} \int \rho\left(\mathbf{r}', t - \frac{R}{v}\right) \frac{d^3 r'}{R}, \quad (8.17a)$$

where integration is extended over all charges of the system under analysis. Acting absolutely similarly, for the vector potential we get

$$\mathbf{A}(\mathbf{r}, t) = \frac{\mu}{4\pi} \int \mathbf{j}\left(\mathbf{r}', t - \frac{R}{v}\right) \frac{d^3 r'}{R}. \quad (8.17b)$$

Retarded
vector
potential

(Now nothing prevents functions $\rho(\mathbf{r}, t)$ and $\mathbf{j}(\mathbf{r}, t)$ from satisfying the continuity relation.)

Solutions (17) are called the *retarded potentials*, the name signifying that the observed fields are “retarded” (delayed) in time by $\Delta t = R/v$ relative to the source variations, due to the finite speed v of the electromagnetic wave propagation. These solutions are so important that they deserve at least a couple of general remarks.

First, remarkably, these simple expressions are *exact* solutions of the Maxwell equations (93) in a uniform medium for an arbitrary distribution of stand-alone charges and currents. They also may be considered as the *general* solutions of these equations, provided that the integration is extended over all field sources in the Universe – or at least in its part that affects our observations.

Second, if functions $\rho(\mathbf{r}, t)$ and $\mathbf{j}(\mathbf{r}, t)$ include the microscopic (bound) charges and currents as well, the macroscopic Maxwell equations (6.93) are valid with the replacement $\varepsilon \rightarrow \varepsilon_0$ and $\mu \rightarrow \mu_0$, so that the retarded potentials solutions (17) are also valid - with the same replacement.

Finally, Eqs. (17) may be plugged into Eqs. (1), giving (after an explicit differentiation) the so-called *Jefimenko equations* for fields \mathbf{E} and \mathbf{B} – similar in structure to Eqs. (17), but more cumbersome. Conceptually, the existence of such equations is a good news, because they are free from the gauge ambiguity pertinent to potentials ϕ and \mathbf{A} . However, the practical value of these explicit expressions for the fields is not too high: for all applications I am aware of, it is easier to use Eqs. (17) to calculate the particular expressions for the potentials first, and only then calculate the fields from Eqs. (1). Let me present the (apparently most important) example of this approach.

8.2. Electric dipole radiation

Consider again the problem that was discussed in electrostatics (Sec. 3.1), namely the field of a localized source with linear dimensions $a \ll r$ (Fig. 1), but now with time-dependent charge and/or current distribution. Using the arguments of that discussion, in particular the condition expressed by Eq. (3.1), $r' \ll r$, we may apply the Taylor expansion (3.3),

$$f(\mathbf{R}) = f(\mathbf{r}) - \mathbf{r}' \cdot \nabla f(\mathbf{r}) + \dots, \quad (8.18)$$

to function $f(\mathbf{R}) \equiv R$ (for which $\nabla f(\mathbf{r}) = \nabla R = \mathbf{n}$, where $\mathbf{n} \equiv \mathbf{r}/r$ is the unit vector directed toward the observation point, see Fig. 1) to approximate distance R as

$$R \approx r - \mathbf{r}' \cdot \mathbf{n}. \quad (8.19)$$

In each of the retarded potential formulas (17), R participates in two places: in the denominator and in the source time argument. If ρ and \mathbf{j} change in time on scale $\sim 1/\omega$, where ω is some characteristic frequency, then any change of argument $(t - R/v)$ on that time scale, for example due to a change of R on the spatial scale $\sim v/\omega = 1/k$, may substantially change these functions. Thus, expansion (18) may be applied to R in the argument $(t - R/v)$ only if $ka \ll 1$, i.e. if the system size a is much smaller than the radiation wavelength $\lambda = 2\pi/k$. On the other hand, function $1/R$ changes relatively slowly, and for it even the first term expansion (19) gives a good approximation as soon as $a \ll r, R$. In this approach, Eq. (17a) yields

$$\phi(\mathbf{r}, t) \approx \frac{1}{4\pi\epsilon r} \int \rho\left(\mathbf{r}', t - \frac{R}{v}\right) d^3 r' = \frac{1}{4\pi\epsilon r} Q\left(t - \frac{R}{v}\right), \quad (8.20)$$

where $Q(t)$ is the net electric charge of the localized system. Due to the charge conservation, this charge cannot change with time, so that the approximation (20) describes just a static Coulomb field of our localized source, rather than a radiated wave.

Let us, however, apply a similar approximation to the vector potential (17b):

$$\mathbf{A}(\mathbf{r}, t) \approx \frac{\mu}{4\pi r} \int \mathbf{j}\left(\mathbf{r}', t - \frac{R}{v}\right) d^3 r'. \quad (8.21)$$

According to Eq. (5.87), in statics the right-hand part of this expression would vanish, but in dynamics this is no longer true. For example, if the current is due to a nonrelativistic motion⁷ of a system of charges q_k , we can write

$$\int \mathbf{j}(\mathbf{r}', t) d^3 r' = \sum_k q_k \dot{\mathbf{r}}_k(t) = \frac{d}{dt} \sum_k q_k \mathbf{r}_k(t) \equiv \dot{\mathbf{p}}(t), \quad (8.22)$$

where $\mathbf{p}(t)$ is the dipole moment of the localized system, defined by Eq. (3.6). Now, after the integration, we may keep only the first term of approximation (19) in the argument $(t - R/v)$ as well, getting

$$\mathbf{A}(\mathbf{r}, t) \approx \frac{\mu}{4\pi r} \dot{\mathbf{p}}\left(t - \frac{r}{v}\right). \quad (8.23)$$

Let us analyze what exactly does this result, valid in the limit $ka \ll 1$, describe. The second of Eqs. (1) allows us to calculate the magnetic field by the spatial differentiation of \mathbf{A} . At large distances $r \gg \lambda$ (i.e. in the so-called *far field zone*), where Eq. (23) describes a virtually plane wave, the main contribution into this derivative is given by the dipole moment factor:

Far
zone
field

$$\mathbf{B}(\mathbf{r}, t) = \frac{\mu}{4\pi r} \nabla \times \dot{\mathbf{p}}\left(t - \frac{r}{v}\right) = -\frac{\mu}{4\pi r v} \mathbf{n} \times \ddot{\mathbf{p}}\left(t - \frac{r}{v}\right). \quad (8.24)$$

This expression means that the magnetic field, at the observation point, is perpendicular to vectors \mathbf{n} and (the retarded value of) $\ddot{\mathbf{p}}$, and its magnitude is

$$B = \frac{\mu}{4\pi r v} \ddot{p}\left(t - \frac{r}{v}\right) \sin \theta, \quad \text{i.e. } H = \frac{1}{4\pi r v} \ddot{p}\left(t - \frac{r}{v}\right) \sin \theta, \quad (8.25)$$

where θ is the angle between those two vectors – see Fig. 2.⁸

⁷ For relativistic particles, moving with velocities of the order of speed of light, one has to be more careful. As the result, I will postpone the discussion of their radiation until Chapter 10, i.e. until after the discussion of special relativity in Chapter 9.

⁸ From the first of Eqs. (1), for the electric field, in the first approximation (23), we would get $-\partial\mathbf{A}/\partial t = -(1/4\pi\epsilon r v) \ddot{\mathbf{p}}(t - r/v) = -(Z/4\pi r) \ddot{\mathbf{p}}(t - r/v)$. The transverse component of this vector (see Fig. 2) is the proper wave field $\mathbf{E} = Z\mathbf{H} \times \mathbf{n}$, while its longitudinal component is exactly compensated by $(-\nabla\phi)$ in the *next* term of expansion of Eq. (17a) with respect to small parameter $r/\lambda \ll 1$.

The most important feature of this result is that the time-dependent field decreases very slowly (only as $1/r$) with the distance from the source, so that the radial component of the corresponding Poynting vector (7.7), $S_r = ZH^2$, drops as $1/r^2$, i.e. the full power \mathcal{P} of the emitted spherical wave, that scales as $r^2 S_r$, does not depend on the distance from the source – as it should for radiation. Equation (25) allows us to be more quantitative; for the instantaneous radiation intensity we may plug it into Eq. (7.9) to get

$$S_r = ZH^2 = \frac{Z}{(4\pi vr)^2} \left[\ddot{\mathbf{p}} \left(t - \frac{r}{v} \right) \right]^2 \sin^2 \theta. \quad (8.26) \quad \text{Instant power density}$$

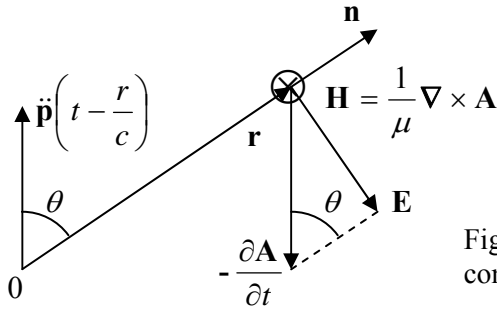


Fig. 8.2. Far zone fields of a localized source, contributing into its electric dipole radiation.

This is the famous formula for the *electric dipole radiation*; this is the dominating component of radiation by a localized system of charges - unless $\ddot{\mathbf{p}} = 0$. Please notice its angular dependence: the radiation vanishes at the axis of the retarded vector $\ddot{\mathbf{p}}$ (where $\theta = 0$), and reaches its maximum in the plane perpendicular to that axis. Integration of S_r over all directions, i.e. over the whole sphere of radius r , gives the total instant power of the dipole radiation:⁹

$$\mathcal{P} \equiv \oint_{r=\text{const}} S_r d^2 r = \frac{Z}{(4\pi v)^2} \ddot{\mathbf{p}}^2 2\pi \int_0^\pi \sin^3 \theta d\theta = \frac{Z}{6\pi v^2} \ddot{\mathbf{p}}^2. \quad (8.27) \quad \text{Full instant power}$$

In order to find the average power, this expression has to be averaged over a sufficiently long time. In particular, if the source is monochromatic, $\mathbf{p}(t) = \text{Re}[\mathbf{p}_\omega \exp\{-i\omega t\}]$, with time-independent vector \mathbf{p}_ω , such averaging may be carried out just over one period, giving an extra factor 2 in the denominator:

$$\overline{\mathcal{P}} = \frac{Z\omega^4}{12\pi v^2} |\mathbf{p}_\omega|^2. \quad (8.28) \quad \text{Full average power}$$

The easiest example of application of the formula is to a point charge oscillating, with frequency ω , along a straight line (that we may take for axis z), with amplitude a . In this case, $\mathbf{p} = q\mathbf{n}_z z(t) = qa \text{Re}[\exp\{-i\omega t\}]$, and if the charge velocity amplitude, $a\omega$, is much less than the wave speed v , we may use Eq. (28) with $p_\omega = qa$, giving

⁹ In the Gaussian units, for free space ($v = c$), this important formula reads $\mathcal{P} = (2/3c^3) \ddot{\mathbf{p}}^2$. It was first derived in 1897 by J. Larmor for the particular case of a single point charge q moving with acceleration $\ddot{\mathbf{r}}$, when $\ddot{\mathbf{p}} = q\ddot{\mathbf{r}}$ and hence $\mathcal{P} = (2q^2/3c^3) \ddot{\mathbf{r}}^2$. As a result, Eq. (27) is sometimes referred to as the *Larmor formula*.

$$\overline{\mathcal{P}} = \frac{Zq^2 a^2 \omega^4}{12\pi v^2}. \quad (8.29)$$

Applied to an electron ($q = -e \approx -1.6 \times 10^{-19}$ C), rotating about a nuclei at an atomic distance $a \sim 10^{-10}$ m, the Larmor formula shows¹⁰ that the energy loss due to the dipole radiation is so large that it would cause electron's collapse on atom's nuclei in just $\sim 10^{-10}$ s. In the beginning of the 1900s, this classical result was one of the main arguments for the development of quantum mechanics that prevents such collapse of electrons in their lowest-energy (ground) state.

Another example of a very useful application of Eq. (28) is the radio wave radiation by a short, straight, symmetric antenna which is fed, for example, by a TEM transmission line such as a coaxial cable – see Fig. 3.

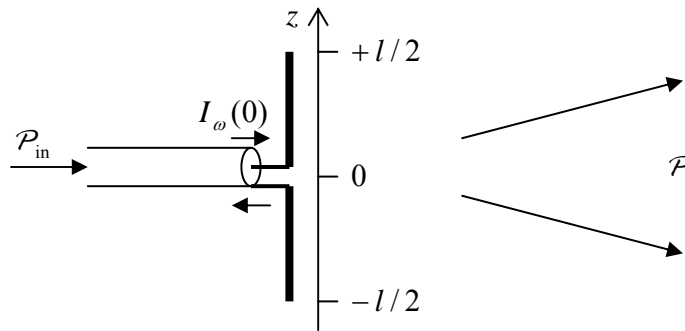


Fig. 8.3. Dipole antenna.

The exact solution of this problem is rather complex, because the law $I_\omega(z)$ of the current variation along antenna's length should be calculated self-consistently with the distribution of the electromagnetic field that is induced by the current in the surrounding space. (This fact is unfortunately ignored in some textbooks.) However, one may argue that at $l \ll \lambda$, the current should be largest in the feeding point (in Fig. 3, taken for $z = 0$), vanish at antenna's ends ($z = \pm l/2$), and that the only possible scale of the current variation in the antenna is l itself, so that the linear function,

$$I_\omega(z) = I_\omega(0) \left(1 - \frac{2}{l}|z|\right), \quad (8.30)$$

gives a good approximation - as it indeed does. Now we can use the continuity equation $\partial Q/\partial t = I$, i.e. $-i\omega Q_\omega = I_\omega$, to calculate the complex amplitude $Q_\omega(z) = iI_\omega(z)\text{sgn}(z)/\omega$ of the electric charge $Q(z, t) = \text{Re}[Q_\omega \exp\{-i\omega t\}]$ of the wire beyond point z , and from it, the amplitude of the linear density of charge

$$\lambda_\omega(z) \equiv \frac{dQ_\omega(z)}{d|z|} = -i \frac{2I_\omega(0)}{\omega l} \text{sgn } z. \quad (8.31)$$

From here, the dipole moment's amplitude is

$$p_\omega = 2 \int_0^{l/2} \lambda_\omega(z) z dz = -i \frac{I_\omega(0)}{2\omega} l, \quad (8.32)$$

¹⁰ Actually, the formula needs a numerical coefficient adjustment to account for electron's orbital (rather than linear) motion – the task left for reader's exercise. However, this adjustment does not affect the order-of-magnitude estimate given above.

so that Eq. (28) yields

$$\overline{\mathcal{P}} = Z \frac{\omega^4}{12\pi v^2} \frac{|I_\omega(0)|^2}{4\omega^2} l^2 = \frac{Z(kl)^2}{24\pi} \frac{|I_\omega(0)|^2}{2}, \quad (8.33)$$

where $k = \omega/v$. The analogy between this result and the dissipation power, $\mathcal{P} = \text{Re}Z (I_\omega^2/2)$, in a lumped linear circuit element, allows the interpretation of the first fraction in the last form of Eq. (33) as the real part of antenna's impedance:

$$\text{Re}Z_A = Z \frac{(kl)^2}{24\pi}, \quad (8.34)$$

as felt by the transmission line. (Indeed, according to Eq. (7.118), the wave traveling along the line toward the antenna is fully radiated, i.e. not reflected back, only if Z_A equals to Z_W of the line.) As we know from Chapter 7, for typical TEM lines, $Z_W \sim Z_0$, while Eq. (34), that is only valid in the limit $kl \ll 1$, shows that for radiation into free space ($Z = Z_0$), $\text{Re}Z_A$ is much less than Z_0 .

Hence in order to reach the impedance matching condition $Z_W = Z_A$, antenna's length should be increased – as a more involved theory shows, to $l \sim \lambda/2$. However, in many cases, practical considerations make short antennas necessary. The most frequently met example nowadays are the cell phone antennas, which use frequencies close to 1 or 2 GHz, with free-space wavelengths λ between 15 and 30 cm, i.e. much larger than the phone size. The quadratic dependence of antenna's efficiency on l , following from Eq. (34), explains why every millimeter counts in the design of such antennas, and why the designs are carefully optimized using software packages for (virtually exact) numerical solution of time-dependent Maxwell equations for the specific shape of the antenna and other phone parts.¹¹

To conclude this section, let me note that if the wave source is not monochromatic, so that $\mathbf{p}(t)$ should be presented as a Fourier series,

$$\mathbf{p}(t) = \text{Re} \sum_{\omega} \mathbf{p}_{\omega} e^{-i\omega t}, \quad (8.35)$$

the terms corresponding to interference of spectral components with different frequencies ω are averaged out at the time averaging of the Poynting vector, so that the *average* radiated power is just a sum of contributions (28) from all substantial frequency components.

8.3. Wave scattering

The formalism described above may be immediately used in the theory of *scattering* – the phenomenon illustrated by Fig. 4. Generally, scattering is a complex problem. However, in many cases it allows the so-called *Born approximation*,¹² in which scattered wave's field applied to the scattering object is assumed to be much weaker than that of the incident wave, and is neglected.

¹¹ A partial list of popular software packages of this kind includes both publicly available codes such as NEC-2 (whose various versions are available online, e.g., at <http://aliath.debian.org/projects/necpp/> and <http://www.qsl.net/4nec2/>), and proprietary packages – such as *Momentum* from Agilent Technologies (now owned by Hewlett-Packard), *FEKO* from EM Software & Systems, and *XFDTD* from Remcom.

¹² Named after M. Born, one of the founding fathers of quantum mechanics. Note, however, the basic idea of this approach was developed much earlier (in 1881) by Lord Rayleigh – see below.

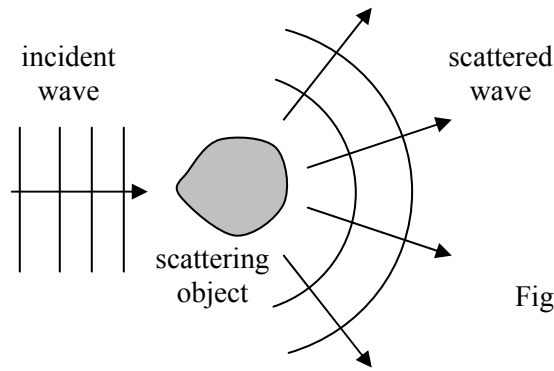


Fig. 8.4. Scattering (schematically).

As the first example of this approach, let us consider scattering of a plane wave, propagating in free space ($Z = Z_0$, $v = c$), by a free¹³ charged particle whose motion may be described by nonrelativistic classical mechanics. (This requires, in particular, the incident wave to be of a modest intensity, so that the speed of the induced charge motion is much less than the speed of light.) In this case the magnetic component of the Lorentz force (5.8),

$$\mathbf{F}_m = q\dot{\mathbf{r}} \times \mathbf{B}, \quad (8.36)$$

exerted on the charge by the magnetic field of a plane wave, is much smaller than force $\mathbf{F}_e = q\mathbf{E}$ exerted by its electric field. Indeed, according to Eq. (7.8), $H = E/Z = E/(\mu/\epsilon)^{1/2}$, $B = \mu H = E/v$, so that the ratio F_m/F_e equals to the ratio of particle's speed, $|\dot{\mathbf{r}}|$, to wave's speed $v \sim c$.

Thus, assuming that the incident wave is linearly-polarized along axis x , the equation of particle's motion in the Born approximation is just $m\ddot{x} = qE(t)$, so that for the x -component $p_x = qx$ of its dipole moment we can write

$$\ddot{p} = q\ddot{x} = \frac{q^2}{m} E(t). \quad (8.37)$$

As we already know from Sec. 2, oscillations of the dipole moment lead to radiation of a wave with a wide angular distribution of intensity; in our case this is the scattered wave – see Fig. 4. Its full power may be found by plugging Eq. (37) into Eq. (27):

$$\mathcal{P} = \frac{Z_0}{6\pi c^2} \ddot{p}^2 = \frac{Z_0 q^4}{6\pi c^2 m^2} E^2(t), \quad \text{i.e. } \bar{\mathcal{P}} = \frac{Z_0 q^4}{12\pi c^2 m^2} |E_\omega|^2. \quad (8.38)$$

Since the power is proportional to incident wave's intensity S , it is customary to characterize scattering ability of the object by the ratio,

$$\sigma \equiv \frac{\bar{\mathcal{P}}}{S_{\text{incident}}} = \frac{\bar{\mathcal{P}}}{|E_\omega|^2 / 2Z_0}, \quad (8.39)$$

which evidently has the dimension of area and is called the *full cross-section* of scattering. For this measure, Eq. (38) yields the famous result

¹³ As Eq. (7.30) shows, this calculation is also valid for an oscillator with eigenfrequency $\omega_0 \ll \omega$.

$$\sigma = \frac{Z_0^2 q^4}{6\pi c^2 m^2} = \frac{\mu_0^2 q^4}{6\pi m^2}, \quad (8.40)$$

which is called the *Thomson scattering formula*,¹⁴ especially when applied to an electron. This relation is most frequently presented in the form¹⁵

$$\sigma = \frac{8\pi}{3} r_c^2, \quad \text{with } r_c \equiv \frac{q^2}{4\pi\epsilon_0} \cdot \frac{1}{mc^2} = 10^{-7} \frac{q^2}{m}. \quad (8.41) \quad \text{Thomson scattering formula}$$

Constant r_c is called the *classical radius of the particle* (or sometimes the “Thomson scattering length”); for electron ($q = -e$, $m = m_e$) it is close to 2.82×10^{-15} m. Its possible interpretation is evident from the first form of Eq. (41) for r_c : at that distance between two similar particles, the potential energy $q^2/4\pi\epsilon_0 r$ of their electrostatic interaction is equal to particle’s rest-mass energy mc^2 .¹⁶

Now we have to go back and establish the conditions at which the Born approximation, when the field of the scattered wave is negligible, is indeed valid for a point-object scattering. Since the scattered wave’s intensity, described by Eq. (26), diverges as $1/r^2$, according to the definition (39) of the cross-section, it may become comparable to S_{incident} at $r^2 \sim \sigma$. However, Eq. (38) itself is only valid if $r \gg \lambda$, so that the Born approximation does not lead to any contradiction if

$$\sigma \ll \lambda^2. \quad (8.42)$$

For the Thompson scattering by an electron, this condition means $\lambda \gg r_c \sim 3 \times 10^{-15}$ m and is fulfilled for all frequencies up to very hard γ rays with energies ~ 100 MeV.

Possibly the most notable feature of result (40) is its independence of the wave frequency. As it follows from its derivation, particularly from Eq. (37), this independence is intimately related with the unbound character of charge motion. For bound charges, say for electrons in a gas molecule, this result is only valid if the wave frequency ω is much higher all eigenfrequencies ω_j of molecular resonances. In the opposite limit, $\omega \ll \omega_j$, the result is dramatically different. Indeed, in this limit we can approximate the molecule’s dipole moment by its static value (3.39)

$$\mathbf{p} = 4\pi\epsilon_0 \alpha_{\text{mol}} \mathbf{E}. \quad (8.43)$$

In the Born approximation, and in the absence of the molecular field effects discussed in Sec. 3.5, \mathbf{E} in this expression is just the incident wave’s field, and we can use Eq. (28) to calculate the power of the wave scattered by a single molecule:

¹⁴ Named after Sir J. J. Thomson (1856-1940), the discoverer of the electron - and isotopes as well! He is not to be confused with his son, G. P. Thomson, who discovered (simultaneously with C. Davisson and L. Germer) quantum-mechanical wave properties of the same electron.

¹⁵ In the Gaussian units, this formula looks like $r_c = q^2/mc^2$ (giving, of course, the same numerical value: for the electron, $r_c \approx 2.82 \times 10^{-13}$ cm). This *classical* quantity should not be confused with particle’s *Compton wavelength* $\lambda_c \equiv h/mc$ (for the electron, close to 2.24×10^{-12} cm), which naturally arises in *quantum* electrodynamics – see a brief discussion in the next chapter, and QM Chapter 9 for more detail.

¹⁶ It is fascinating how smartly has the *relativistic* expression mc^2 sneaked into the result (40), which was obtained using a *nonrelativistic* equation of particle motion. This was possible because the calculation engaged electromagnetic waves that propagate with the speed of light, and whose quanta (*photons*), as a result, may be frequently treated as relativistic (moreover, ultrarelativistic) particles - see the next chapter.

$$\overline{\mathcal{P}} = \frac{4\pi Z_0 \omega^4 \varepsilon_0^2}{c^2} \alpha_{\text{mol}}^2 |E_\omega|^2. \quad (8.44)$$

Now, using the last form of definition (39) of the cross-section, we get a very simple result,

$$\sigma = \frac{8\pi Z_0^2 \varepsilon_0^2 \omega^4}{3c^2} \alpha_{\text{mol}}^2 = \frac{8\pi k^4}{3} \alpha_{\text{mol}}^2, \quad (8.45)$$

showing that in contrast to Eq. (40), at low frequencies σ grows as fast as ω^4 .

Now let us explore the effect of such *Rayleigh scattering*¹⁷ on wave propagation in a gas, with relatively low density n . We can expect (and will prove in the next section) that due to the randomness of molecule positions, the waves scattered by each molecules may be treated as *incoherent*, so that the total scattering power may be calculated just as the sum of those scattered by each molecule. We can use this additivity to write the balance of the incident's wave intensity on a small volume dV of length (along the incident wave direction) dz , and area A in across it. Since such a segment includes $ndV = nAdz$ molecules, and, according to definition (39), each of them scatters power $S\sigma = \mathcal{P}\sigma/A$, the total scattered power is $n\mathcal{P}\sigma dz$; hence the incident power's change is

$$d\mathcal{P} \equiv -n\sigma\mathcal{P} dz. \quad (8.46)$$

Comparing this equation with the general definition (7.202) of the attenuation constant, we see that scattering gives the following contribution to attenuation: $\alpha = n\sigma$. From here, using Eq. (3.41) to write $\alpha_{\text{mol}} = (\varepsilon_r - 1)/4\pi n$, and Eq. (45), we get

$$\alpha = \frac{k^4}{6\pi n} (\varepsilon_r - 1)^2. \quad (8.47)$$

Rayleigh
scattering
formula

This is the famous *Rayleigh scattering formula*, which in particular explains the colors of blue sky and red sunsets. Indeed, through the visible light spectrum, ω changes almost two-fold; as a result, scattering of blue components of sunlight is an order of magnitude higher than that of its red components. More qualitatively, for air near the Earth surface, $\varepsilon_r - 1 \approx 6 \times 10^{-4}$, and $n \sim 2.5 \times 10^{25} \text{ m}^{-3}$ - see Sec. 3.3. Plugging these numbers into Eq. (47), we see that the characteristic length $\mathcal{L} \equiv 1/\alpha$ of scattering is $\sim 30 \text{ km}$ for blue light and $\sim 200 \text{ km}$ for red light.¹⁸ The Earth atmosphere is thinner ($h \sim 10 \text{ km}$), so that the Sun looks just a bit yellowish during most of the day. However, elementary geometry shows that on sunset, the light should pass length $l \sim (R_E h)^{1/2} \approx 300 \text{ km}$ to reach an Earth-surface observer; as a result, the blue components of Sun's light spectrum are almost completely scattered out, and even the red components are weakened considerably.

To conclude the discussion of Eq. (47), let me note that its comparison with the condition of the direct applicability of the Born approximation for a distributed object of size a :

$$\alpha a \ll 1, \quad (8.48)$$

¹⁷ Named after Lord Rayleigh (born J. Stoff, 1842-1919), whose numerous contributions to science include the discovery of argon. He has also pioneered (for the special case we are considering now) the basic idea of what is presently called the Born approximation.

¹⁸ These values are approximate because both n and $(\varepsilon_r - 1)$ vary through the atmosphere.

implies, in particular, that if the electric polarizability of the material is small, $\varepsilon_r \rightarrow 1$, we may be able to use the approximation for an analysis of scattering by even relatively large objects, with size of the order of, or even larger than λ . However, for such extended objects, the phase difference factors (neglected above) step in, leading in particular to the important effects of *interference* and *diffraction*, to whose discussion we now proceed.

8.4. Interference and diffraction

These effects show up not as much in the total power of scattered radiation, as in its angular distribution. It is traditional to characterize this distribution by the *differential cross-section* defined as

$$\frac{d\sigma}{d\Omega} \equiv \frac{\overline{S}_r r^2}{S_{\text{incident}}}, \quad (8.49)$$

Differential cross-section: definition

where r is the distance from the scatterer, at which the scattered wave is observed. Both the definition and notation become more clear if we notice that according to Eq. (26), at large distances ($r \gg a$), the nominator in the right-hand part of Eq. (49), and hence the differential cross-section as the whole, does not depend on r , and that its integral over the total solid angle $\Omega = 4\pi$ coincides with the total cross-section defined by Eq. (39):

$$\oint_{4\pi} \frac{d\sigma}{d\Omega} d\Omega = \frac{1}{S_{\text{incident}}} r^2 \oint_{4\pi} \overline{S}_r d\Omega = \frac{1}{S_{\text{incident}}} \oint_{r=\text{const}} \overline{S}_r d^2r = \frac{\overline{\mathcal{P}}}{S_{\text{incident}}} \equiv \sigma. \quad (8.50)$$

For example, according to Eq. (26), the angular distribution of radiation scattered by a point linear dipole, in the Born approximation, is rather broad; in particular, in the low-frequency limit (43),

$$\frac{d\sigma}{d\Omega} = k^4 \alpha_{\text{mol}}^2 \sin^2 \theta. \quad (8.51)$$

If the wave is scattered by a small dielectric body, with a characteristic size $a \ll \lambda$ (i.e., $ka \ll 1$), then all its parts re-radiate the incident wave coherently. Hence, we can calculate it in the similar way, just replacing the molecular dipole moment (43) with the total dipole moment of the object – see Eq. (3.37):

$$\mathbf{p} = \mathbf{P}V = (\varepsilon_r - 1)\varepsilon_0 \mathbf{E}V, \quad (8.52)$$

where $V \sim a^3$ is body's volume. As a result, the differential cross-section may be obtained from Eq. (51) with the replacement $\alpha_{\text{mol}} \rightarrow V(\varepsilon_r - 1)/4\pi$.

$$\frac{d\sigma}{d\Omega} = \frac{k^4 V^2}{(4\pi)^2} (\varepsilon_r - 1)^2 \sin^2 \theta, \quad (8.53)$$

i.e. follows the same $\sin^2 \theta$ law. The situation for extended objects, with at least one dimension of the order, or larger than the wavelength, is different: here we have to take into account that the phase shifts introduced by various parts of the body are different. Let us analyze this issue for an arbitrary collection of similar point scatterers located at points \mathbf{r}_j .

If wave vector of the incident plane wave is \mathbf{k}_0 , the field the wave has the phase factor $\exp\{i\mathbf{k}_0 \cdot \mathbf{r}\}$ – see Eq. (7.79). At the location of j -th scattering center, the factor equals to $\exp\{i\mathbf{k}_0 \cdot \mathbf{r}_j\}$, so that the local polarization vector \mathbf{p} , and the scattered wave it creates, are proportional to this factor. On

its way to the observation point \mathbf{r} , the scattered wave, with wave vector \mathbf{k} (with $k = k_0$), acquires an additional phase factor $\exp\{i\mathbf{k} \cdot (\mathbf{r} - \mathbf{r}_j)\}$, so that the scattered wave field is proportional to

$$\exp\{i\mathbf{k}_0 \cdot \mathbf{r}_j + i\mathbf{k}(\mathbf{r} - \mathbf{r}_j)\} = \exp\{i(\mathbf{k}_0 - \mathbf{k}) \cdot \mathbf{r}_j + i\mathbf{k} \cdot \mathbf{r}\} = e^{i\mathbf{k} \cdot \mathbf{r}} \exp\{-i(\mathbf{k} - \mathbf{k}_0) \cdot \mathbf{r}_j\}. \quad (8.54)$$

Since the first factor in the last expression does not depend on \mathbf{r}_j , in order to calculate the total scattering wave, it is sufficient to sum up the elementary phase factors $\exp\{-i\mathbf{q} \cdot \mathbf{r}_j\}$, where vector

$$\mathbf{q} \equiv \mathbf{k} - \mathbf{k}_0 \quad (8.55)$$

has the physical sense of the wave vector change at scattering.¹⁹ It may look like the phase factor depends on the choice of origin. However, according to Eq. (7.42), the average intensity of the scattered wave is proportional to $E_\omega E_\omega^*$, i.e. to the following real scalar function of vector \mathbf{q} :

Scattering
function

$$F(\mathbf{q}) = \left(\sum_j \exp\{-i\mathbf{q} \cdot \mathbf{r}_j\} \right) \left(\sum_{j'} \exp\{-i\mathbf{q} \cdot \mathbf{r}_{j'}\} \right)^* = \sum_{j,j'} \exp\{i\mathbf{q} \cdot (\mathbf{r}_j - \mathbf{r}_{j'})\} \equiv |I(\mathbf{q})|^2, \quad (8.56)$$

where the complex function

Phase
sum

$$I(\mathbf{q}) \equiv \sum_j \exp\{-i\mathbf{q} \cdot \mathbf{r}_j\} \quad (8.57)$$

is called the *phase sum*, may be calculated within any reference frame, without affecting the final result (56). The double-sum form of Eq. (56) is convenient to notice that for a system of *many* ($N \gg 1$) of similar but randomly located scatterers, only the terms with $j = j'$ accumulate at summation, so that $F(\mathbf{q})$ scales as N , rather than N^2 - thus justifying the above treatment of the Rayleigh scattering problem.

Let us start using Eq. (56) by applying it to the simplest problem of just *two* similar small scatterers, separated by a fixed distance a :

$$F(\mathbf{q}) = \sum_{j,j'=1}^2 \exp\{i\mathbf{q} \cdot (\mathbf{r}_j - \mathbf{r}_{j'})\} = 2 + \exp\{-iq_a a\} + \exp\{iq_a a\} = 2(1 + \cos q_a a) = 4 \cos^2 \frac{q_a a}{2}, \quad (8.58)$$

where $q_a \equiv \mathbf{q} \cdot \mathbf{a}/a$ is the component of vector \mathbf{q} along vector \mathbf{a} connecting the scatterers. The apparent simplicity of this result may be a bit misleading, because the mutual plane of vectors \mathbf{k} and \mathbf{k}_0 (and hence of vector \mathbf{q}) does not necessarily coincide with the mutual plane of vectors \mathbf{k}_0 and \mathbf{E}_ω , so that the *scattering angle* α between vectors \mathbf{k} and \mathbf{k}_0 is generally different from $(\pi/2 - \theta)$ - see Fig. 5.

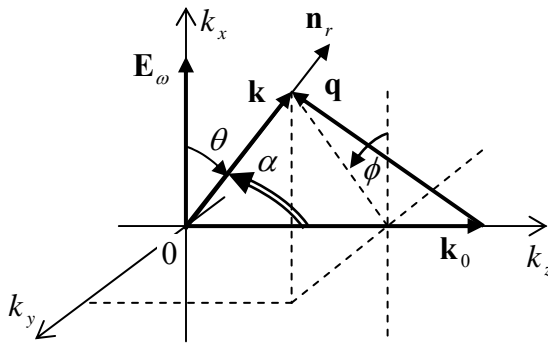


Fig. 8.5. Angles important for the general scattering problem.

¹⁹ In quantum electrodynamics, $\hbar\mathbf{q}$ has the sense of the momentum transferred from the scattering object to the scattered photon, and this terminology sometimes creeps even into the classical electrodynamic texts.

Moreover, vectors \mathbf{q} and \mathbf{a} may have another common plane, and angle between them is one more parameter that may be considered as independent from both α and θ . As a result, the angular dependence of the scattered wave's intensity (and hence $d\sigma/d\Omega$), that depends on all three angles, may be rather complex.

This is why let me consider only the simple case when vectors \mathbf{k} , \mathbf{k}_0 , and \mathbf{a} are all in the same plane (Fig. 6a), with \mathbf{k}_0 perpendicular to \mathbf{a} (leaving the general analysis for readers' exercise). Then, with our choice of coordinates, $q_a = q_x = k \sin \alpha$, and Eq. (58) is reduced to

$$F(\mathbf{q}) = 4 \cos^2 \frac{ka \sin \alpha}{2}. \quad (8.59)$$

This function always has two maxima, at $\alpha = 0$ and $\alpha = \pi$, and possibly (if the product ka is large enough) other maxima at special angles α_n that satisfy the famous *Bragg condition*²⁰

$$ka \sin \alpha_n = 2\pi n, \quad \text{i.e. } a \sin \alpha_n = n\lambda. \quad (8.60) \quad \text{Bragg condition}$$

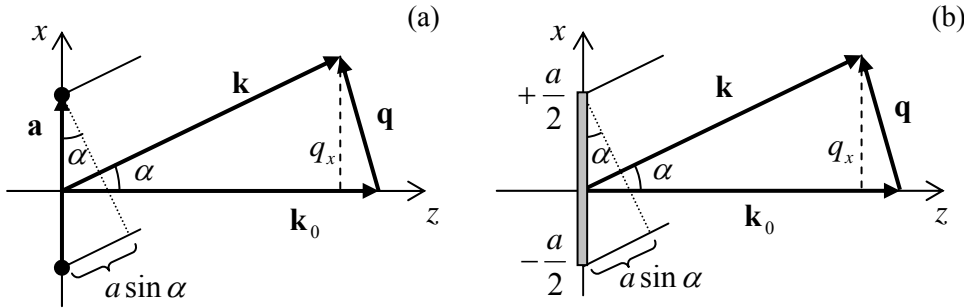


Fig. 8.6. The simplest geometries for (a) interference and (b) diffraction.

As evident from Fig. 6a, this condition may be readily understood as the in-phase addition (frequently called the *constructive interference*) of two coherent waves scattered by the two points, when the difference between their paths toward the observer, $a \sin \alpha$, equals to an integer number of wavelengths. At each such maximum, $F = 4$, due to the doubling of the wave amplitude and hence quadrupling its power.

If the distance between the point scatterers is large ($ka \gg 1$), the first Bragg maxima correspond to small angles, $\alpha \ll 1$. For this region, Eq. (59) is reduced to a simple sinusoidal dependence of function F on angle α . Moreover, within the range of small α , the polarization factor $\sin^2 \theta$ is virtually constant, so that the scattered wave intensity, and hence the differential cross-section

$$\frac{d\sigma}{d\Omega} \propto F(\mathbf{q}) = 4 \cos^2 \frac{ka\alpha}{2}. \quad (8.61) \quad \text{Young's interference pattern}$$

This is of course the well-known *interference pattern*, well known from the Young's two-slit experiment.²¹ (As will be discussed in the next section, theoretical description of the two-slit experiment

²⁰ Named after Sir William Bragg and his son, Sir William Lawrence Bragg, who in 1912 demonstrated X-ray diffraction by atoms in crystals. The Braggs' experiments have made the existence of atoms (before that, a hypothetical notion ignored by many physicists) indisputable.

is more complex than that of the Born scattering, but is preferable experimentally, because at scattering, the wave of intensity (61) has to be observed on the backdrop of a stronger incident wave that propagates in almost the same direction, $\alpha = 0$.)

The Bragg condition (60) does not change at scattering from $N > 2$ similar, equidistant scatterers, located along the same straight line (because the condition is applicable to each pair of adjacent scatterers), but the interference pattern changes. Leaving the analysis of the case of arbitrary N for reader's exercise, let me jump to the limit $N \rightarrow 0$, in which we may ignore the scatterer discreteness. The resulting pattern is similar to that at scattering by a continuous thin rod, so let us first discuss the Born scattering by an arbitrary distributed object - say an extended dielectric body with a constant value of ε_r . Transferring Eq. (56) from the sum to an integral, for the differential cross-section we get

$$\frac{d\sigma}{d\Omega} = \frac{k^4}{(4\pi)^2} (\varepsilon_r - 1)^2 F(\mathbf{q}) \sin^2 \theta = \frac{k^4}{(4\pi)^2} (\varepsilon_r - 1)^2 |I(\mathbf{q})|^2 \sin^2 \theta, \quad (8.62)$$

where $I(\mathbf{q})$ now becomes the *phase integral*,²²

Phase
integral

$$I(\mathbf{q}) = \int_V \exp\{-i\mathbf{q} \cdot \mathbf{r}'\} d^3 r', \quad (8.63)$$

with the dimensionality of volume.

Now we may return to the particular case of a thin rod (with both dimensions of the cross-section's area much smaller than λ , but an arbitrary length a), otherwise keeping the same simple geometry as for two point scatterers – see Fig. 6b. In this case the phase integral is just

Fraunhofer
diffraction
integral

$$I(\mathbf{q}) = A \int_{-a/2}^{+a/2} \exp\{-iq_x x'\} dx' = A \frac{\exp\{-iq_x a/2\} - \exp\{iq_x a/2\}}{-iq_x} = V \frac{\sin \xi}{\xi}, \quad (8.64)$$

where $V = Aa$ is the volume of the rod, and ξ is a dimensionless parameter defined as

$$\xi \equiv \frac{q_x a}{2} = \frac{ka \sin \alpha}{2}. \quad (8.65)$$

The fraction participating in Eq. (64) is met in physics so frequently that it has deserved the special name *sinc* (not “sync”, please!) *function*:

Sinc
function

$$\text{sinc} \xi \equiv \frac{\sin \xi}{\xi}. \quad (8.66)$$

Obviously, this function, plotted in Fig. 7, vanishes at all points $\xi_n = \pi n$, with integer n , besides point $n = 0$: $\text{sinc} \xi_0 = \text{sinc} 0 = 1$.

²¹ This experiment was described as early as in 1803 by T. Young – one more universal genius of science, who has also introduced the Young modulus in the elasticity theory (see, e.g., CM Chapter 7), besides numerous other achievements - including deciphering Egyptian hieroglyphs! The two-slit experiment has firmly established the wave picture of light, to be replaced by the dualistic photon-vs-wave picture, formalized by quantum electrodynamics, only 100+ years later.

²² Since the observation point's position \mathbf{r} does not participate in this formula explicitly, the prime sign in \mathbf{r}' could be dropped, but I keep it as a reminder that the integral is taken over points \mathbf{r}' of the *scattering object*.

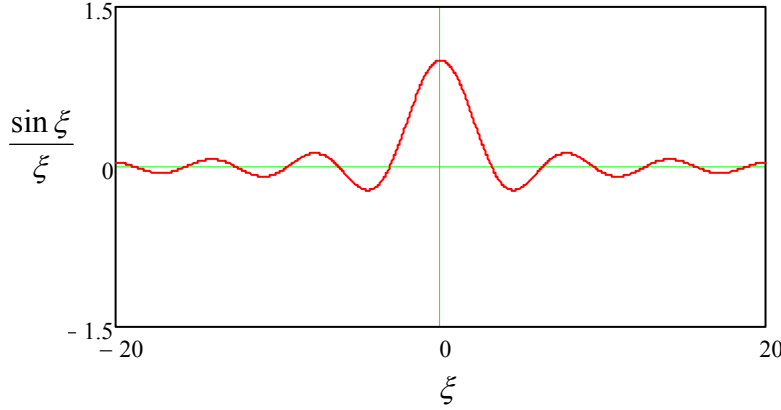


Fig. 8.7. Sinc function.

The function $F(\mathbf{q}) = V^2 \text{sinc}^2 \xi$, resulting from Eq. (64), is plotted by red line in Fig. 8, and is called the *Fraunhofer diffraction pattern*.

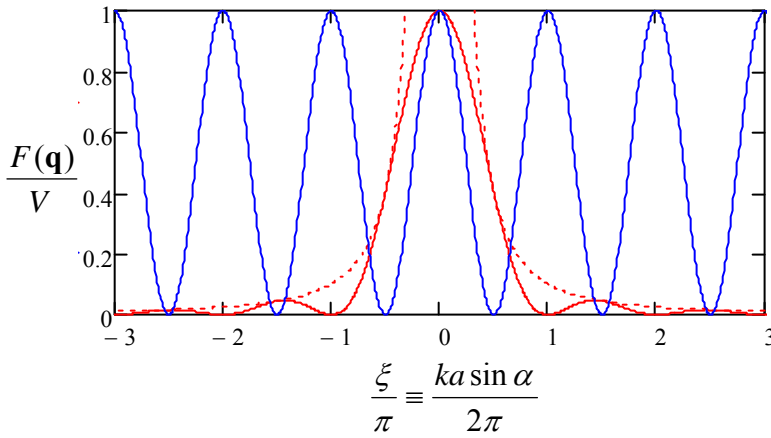


Fig. 8.8. The Fraunhofer diffraction pattern (solid red line) and its envelope $1/\xi^2$ (dashed line). For comparison, the blue line shows the standard interference pattern $\cos^2 \xi$ - cf. Eq. (59).

Note that it oscillates with the same argument period $\Delta(ka \sin \alpha) = 2\pi/ka \ll 1$ as the interference pattern (59) from two point scatterers (shown with the blue line in Fig. 8). However, at the interference, the scattered wave intensity vanishes at angles α_n that satisfy condition

$$\frac{ka \sin \alpha'_n}{2\pi} = n + \frac{1}{2}, \quad (8.67)$$

when the optical paths difference $a \sin \alpha$ equals to a semi-integer number of wavelengths $\lambda/2 = \pi/k$, and hence the two waves from the scatterers arrive to the observer in anti-phase (the so-called *destructive interference*). On the other hand, for the diffraction from a continuous rod the minima occur at a different set of angles,

$$\frac{ka \sin \alpha_n}{2\pi} = n, \quad (8.68)$$

i.e. exactly where the two-point interference pattern has its maxima. The reason for this relation is that the wave diffraction on the rod may be considered as a simultaneous interference of waves from all its fragments, and exactly at the observation angles when the rod edges give waves with phases shifted by $2\pi n$, the interior point of the rod give waves with all possible phases, with their algebraic sum equal to

zero. Even more visibly in Fig. 8, at diffraction the intensity oscillations are limited by a rapidly decreasing envelope function $1/\xi^2$. The reason for this fast decrease is that with each Fraunhofer diffraction period, a smaller and smaller fraction of the rod gives an unbalanced contribution to the scattered wave.

If rod's length is small ($ka \ll 1$, i.e. $a \ll \lambda$), then sinc's argument ξ is small at all scattering angles α , so $I(\mathbf{q}) \approx V$, and Eq. (64) is reduced to Eq. (53). In the opposite limit, $a \gg \lambda$, the first zeros of function $I(\mathbf{q})$ correspond to very small angles α , for which $\sin\theta \approx 1$, so that the differential cross-section is

$$\frac{d\sigma}{d\Omega} = \frac{k^4 V^2}{(4\pi)^2} (\epsilon_r - 1)^2 \text{sinc}^2 \frac{ka\alpha}{2}, \quad (8.69)$$

i.e. Fig. 8 shows the scattering intensity as a function of the diffraction direction – if the pattern is observed within the plane containing the rod.

8.5. The Huygens principle

The Born approximation allows tracing the basic features of (and the difference between) the phenomena of interference and diffraction. Unfortunately, this approximation, based on the relative weakness of the scattered wave, cannot be used for more popular experimental implementations of these phenomena, for example, the Young's two-slit experiment, or diffraction on a single slit or orifice – see, e.g. Fig. 9. Indeed, at such experiments, the orifice size a is typically much larger than light's wavelength, and as a result, no clear decomposition of the fields to the incident and “scattered” waves is possible.

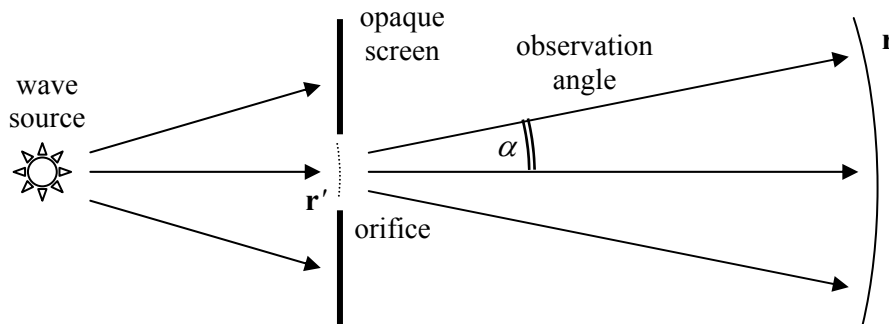


Fig. 8.9. Typical geometry for the Huygens principle application.

However, for such experiments, another approximation, called the *Huygens* (or “Huygens-Fresnel”) *principle*,²³ is very instrumental: the passed wave may be presented as a linear superposition of spherical waves of the type (17), as if they were emitted by every point of the orifice (or more physically, by every point of the incident wave's front that has arrived at the orifice). This approximation is valid if the following strong conditions are satisfied:

²³ Named after C. Huygens (1629-1695) who had conjectured the wave theory of light (that remained controversial for more than a century, until T. Young's experiments), and A.-J. Fresnel (1788-1827) who has developed the mathematical theory of diffraction.

$$\lambda \ll a \ll r, \quad (8.70)$$

where r is the distance of the observation point from the orifice. In addition, as we have seen in the last section, at small λ/a the diffraction phenomena are confined to angles $\alpha \sim 1/ka \sim \lambda/a \ll 1$. For observation at such small angles, the mathematical expression of the Huygens principle, for a complex amplitude $f_\omega(\mathbf{r})$ of a monochromatic wave $f(\mathbf{r}, t) = \text{Re}[f_\omega e^{-i\omega t}]$, is given by the following simple formula

$$f_\omega(\mathbf{r}) = C \int_{\text{orifice}} f_\omega(\mathbf{r}') \frac{e^{ikR}}{R} d^2 r'. \quad (8.71)$$

Here f is any transverse component of any of wave's fields (either \mathbf{E} or \mathbf{H}),²⁴ R is the distance between point \mathbf{r}' at the orifice and the observation point \mathbf{r} (i.e. the magnitude of vector $\mathbf{R} \equiv \mathbf{r} - \mathbf{r}'$), and C is a complex constant.

Before describing the proof of Eq. (71), let me carry out its sanity check - which also will give us the constant C . Let us see what happens if the field under the integral is the usual plane wave $f_\omega(z)$ propagating along axis z (i.e. there is no opaque screen at all), so we should take the whole x - y plane, say with $z' = 0$, as the integration area (Fig. 10).

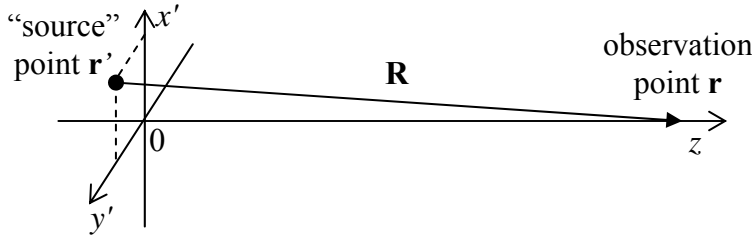


Fig. 8.10. The Huygens principle applied to a plane wave.

Then, for the observation point with coordinates $x = 0$, $y = 0$, and $z \gg \lambda$, Eq. (71) yields

$$f_\omega(z) = C f_\omega(0) \int dx' \int dy' \frac{\exp\{ik(x'^2 + y'^2 + z^2)^{1/2}\}}{(x'^2 + y'^2 + z^2)^{1/2}}. \quad (8.72)$$

Before specifying the integration limits, let us consider the range $|x'|, |y'| \ll z$. In this range the square root, met in Eq. (72) twice, may be approximated as

$$(x'^2 + y'^2 + z^2)^{1/2} = z \left(1 + \frac{x'^2 + y'^2}{z^2} \right)^{1/2} \approx z \left(1 + \frac{x'^2 + y'^2}{2z^2} \right) = z + \frac{x'^2 + y'^2}{2z}. \quad (8.73)$$

The denominator of Eq. (72) is a much slower function of x' and y' than the exponent, and in it (as we will check *a posteriori*), it is sufficient to keep just the main, first term of expansion (73). With that, Eq. (72) becomes

²⁴ The fact that the Huygens principle is valid for any field component should not too surprising. Due to condition $a \gg \lambda$, the real boundary conditions at the orifice edges are not important; what is only important that the screen, that limits the orifice, is opaque. Because of this, the Huygens principle's expression (71) is a part of the so-called *scalar theory of diffraction*. (In this course I will not have time to go beyond this approximation.)

$$f_{\omega}(z) = Cf_{\omega}(0) \frac{e^{ikz}}{z} \int dx' \int dy' \exp \frac{ik(x'^2 + y'^2)}{2z} = Cf_{\omega}(0) \frac{e^{ikz}}{z} I_x I_y, \quad (8.74)$$

where I_x and I_y are two similar integrals; for example,

$$I_x = \int \exp \frac{ikx'^2}{2z} dx' = \left(\frac{2z}{k} \right)^{1/2} \int \exp \{i\xi^2\} d\xi = \left(\frac{2z}{k} \right)^{1/2} \left[\int \cos(\xi^2) d\xi + i \int \sin(\xi^2) d\xi \right], \quad (8.75)$$

where $\xi \equiv (k/2z)^{1/2}$. These are the so-called *Fresnel integrals*. I will discuss them in more detail in the next section, and right now, only one property of these integrals is important for us: if taken in symmetric limits $[-\xi_0, +\xi_0]$, both of them rapidly converge to the same value, $(\pi/2)^{1/2}$, as soon as ξ_0 becomes much larger than 1.²⁵ This means that even if we do not impose any exact limits on the integration area in Eq. (72), this integral converges to value

$$f_{\omega}(z) = Cf_{\omega}(0) \frac{e^{ikz}}{z} \left\{ \left(\frac{2z}{k} \right)^{1/2} \left[\left(\frac{\pi}{2} \right)^{1/2} + i \left(\frac{\pi}{2} \right)^{1/2} \right] \right\}^2 = \left(C \frac{2\pi i}{k} \right) f_{\omega}(0) e^{ikz}, \quad (8.76)$$

due to contributions from the central area with linear size of the order of $\Delta\xi \sim 1$, i.e.

$$\Delta x \sim \Delta y \sim \left(\frac{z}{k} \right)^{1/2} \sim (\lambda z)^{1/2}, \quad (8.77)$$

so that the contribution by front points \mathbf{r}' well beyond the range (77) is negligible.²⁶ (Within our assumptions (70), which in particular require λ to be much less than z , the *diffraction angle* $\Delta x/z \sim \Delta y/z \sim (\lambda/z)^{1/2}$, corresponding to the important area of the front, is small.) In order to sustain the plane wave propagation, $f_{\omega}(z) = f_{\omega}(0)e^{ikz}$, constant C in Eq. (76) has to be taken equal to $k/2\pi i$. Thus, the Huygens principle's prediction (71), in its final form, reads

$$f_{\omega}(\mathbf{r}) = \frac{k}{2\pi i} \int_{\text{orifice}} f_{\omega}(\mathbf{r}') \frac{e^{ikR}}{R} d^2 r', \quad (8.78)$$

and describes, in particular, the straight propagation of the plane wave (in a uniform media).

Let me pause to emphasize how nontrivial this result is. It would be a natural corollary of Eq. (25) (and the linear superposition principle) if all points of the orifice were filled with point scatterers that re-emit all the incident waves into spherical waves. However, as it follows from the above proof, the Huygens principle is also valid if there is nothing in the orifice but the free space!

This is why it is important a proof of the principle,²⁷ based on the Green's theorem (2.207). Let us apply this theorem to function $f = f_{\omega}$, where f_{ω} is the complex amplitude of a scalar component of one of wave's fields, which satisfies the Helmholtz equation (7.192),

²⁵ See, e.g., MA Eq. (6.10).

²⁶ This result very is natural, because $\exp\{ikR\}$ oscillates fast with the change of \mathbf{r}' , so that the contributions from various front point are averaged out. Indeed, the only reason why the central part of plane $[x', y']$ gives a nonvanishing contribution (76) to $f_{\omega}(z)$ is that the phase exponents stops oscillating at $(x'^2 + y'^2)$ below $\sim z/k$ – see Eq. (73).

²⁷ This proof was given in 1882 by G. Kirchhoff.

$$(\nabla^2 + k^2)f_\omega(\mathbf{r}) = 0, \quad (8.79)$$

and function $g = g_\omega$, which is the time Fourier image of the corresponding Green's function. It may be defined, as usual, as the solution to the same equation with the added delta-functional right-hand part with an arbitrary coefficient, for example,

$$(\nabla^2 + k^2)g_\omega(\mathbf{r}, \mathbf{r}') = -4\pi\delta(\mathbf{r} - \mathbf{r}'). \quad (8.80)$$

With Eqs. (79) and (80) used to express the Laplace operators of functions f_ω and g_ω , Eq. (2.207) becomes

$$\int_V \left\{ f_\omega \left[-k^2 g_\omega(\mathbf{r}, \mathbf{r}') - 4\pi\delta(\mathbf{r} - \mathbf{r}') \right] - g_\omega(\mathbf{r}, \mathbf{r}') \left[-k^2 f_\omega \right] \right\} d^3r = \oint_S \left[f_\omega \frac{\partial g_\omega(\mathbf{r}, \mathbf{r}')}{\partial n} - g_\omega(\mathbf{r}, \mathbf{r}') \frac{\partial f_\omega}{\partial n} \right] d^2r, \quad (8.81)$$

where \mathbf{n} is the outward normal to the surface S limiting volume V . Two terms in the left-hand side of this relation cancel, so that after swapping \mathbf{r} and \mathbf{r}' we get

$$-4\pi f_\omega(\mathbf{r}) = \oint_S \left[f_\omega(\mathbf{r}') \frac{\partial g_\omega(\mathbf{r}', \mathbf{r})}{\partial n'} - g_\omega(\mathbf{r}', \mathbf{r}) \frac{\partial f_\omega(\mathbf{r}')}{\partial n'} \right] d^2r'. \quad (8.82)$$

This relation is only correct if the selected volume V includes point \mathbf{r} (otherwise we would not get its left-hand part from the integration of the delta-function), but does not include the genuine source of the wave (otherwise Eq. (79) would have a nonvanishing right-hand part). Let \mathbf{r} be the field observation point, V all the source-free half-space (for example, the half-space right of the screen in Fig. 9), so that S is the surface of the screen, including the orifice. Then the right-hand part of Eq. (82) describes the field in the observation point \mathbf{r} induced by the wave passing through the orifice points \mathbf{r}' . Since no waves are emitted by the opaque parts of the screen, we can limit the integration by the orifice area.²⁸ Assuming also that the opaque parts of the screen do not re-emit waves “radiated” by the orifice, we can take the solution of Eq. (80) to be the retarded potential for the free space:²⁹

$$g_\omega(\mathbf{r}, \mathbf{r}') = \frac{e^{ikR}}{R}. \quad (8.83)$$

Plugging this expression into Eq. (82), we get

$$-4\pi f_\omega(\mathbf{r}) = \oint_{\text{orifice}} \left[f_\omega(\mathbf{r}') \frac{\partial}{\partial n'} \left(\frac{e^{ikR}}{R} \right) - \left(\frac{e^{ikR}}{R} \right) \frac{\partial f_\omega(\mathbf{r}')}{\partial n'} \right] d^2r'. \quad (8.84) \quad \text{Kirchhoff integral}$$

This is the so-called *Kirchhoff* (or “Fresnel-Kirchhoff”) *integral*.³⁰ Now, let us make the two additional approximations. The first of them stems from Eq. (70): at $ka \gg 1$, the wave's spatial dependence in the orifice area may be presented as

²⁸ Actually, this is a somewhat nontrivial point of the proof. Indeed, it may be shown that the solution of Eq. (79) identically equals to zero if $f(\mathbf{r}')$ and $\partial f(\mathbf{r}')/\partial n'$ vanish together at any part of the boundary. As a result, building the solution with the account of exact boundary conditions (which is the task of the vector theory of diffraction) is possible but cumbersome. Here we base our solution on the physical intuition.

²⁹ It follows, e.g., from Eq. (16) with a monochromatic source $q(t) = q_0 \exp\{-i\omega t\}$, at the value $q_\omega = 4\pi\epsilon$ that fits the right-hand part of Eq. (80).

³⁰ With the integration extended over *all* boundaries of volume V , this would be an exact result.

$$f_{\omega}(\mathbf{r}') = (\text{a slow function of } \mathbf{r}') \times \exp\{i\mathbf{k}_0 \cdot \mathbf{r}'\}, \quad (8.85)$$

where “slow” means a function that changes on the scale of a rather than λ . If, also, $kR \gg 1$, then the differentiation in Eq. (84) may be, in both instances, limited to the rapidly changing exponents, giving

$$-4\pi f_{\omega}(\mathbf{r}) = \oint_A i(\mathbf{k} + \mathbf{k}_0) \cdot \mathbf{n}' \frac{e^{ikR}}{R} f(\mathbf{r}') d^2 r', \quad (8.86)$$

Second, if all observation angles are small, we can take $\mathbf{k} \cdot \mathbf{n}' \approx \mathbf{k}_0 \cdot \mathbf{n}' \approx -k$. With that, Eq. (86) is reduced to Eq. (78) expressing the Huygens principle.

It is clear that the principle immediately gives a very simple description of the interference of waves passing through two small holes in the screen. Indeed, if the hole size is negligible in comparison with distance a between them (though still much larger than the wavelength!), Eq. (78) yields

$$f_{\omega}(\mathbf{r}) = c_1 e^{ikR_1} + c_2 e^{ikR_2}, \quad \text{with } c_{1,2} \equiv \frac{k f_{1,2} A_{1,2}}{2\pi i R_{1,2}}, \quad (8.87)$$

where $R_{1,2}$ are the distances between the holes and the observation point, and $A_{1,2}$ are the hole areas. For the interference wave intensity, Eq. (87) yields

$$\bar{S} \propto f_{\omega} f_{\omega}^* = |c_1|^2 + |c_2|^2 + 2|c_1||c_2|\cos[k(R_1 - R_2) + \varphi], \quad \varphi \equiv \arg c_1 - \arg c_2. \quad (8.88)$$

The first two terms in this result clearly represent the intensities of partial waves passed through each hole, while the last one the result of their interference. The interference pattern's *contrast ratio*

$$\mathcal{R} \equiv \frac{\bar{S}_{\max}}{\bar{S}_{\min}} = \left(\frac{|c_1| + |c_2|}{|c_1| - |c_2|} \right)^2, \quad (8.89)$$

is largest (infinite) when both waves have equal amplitudes.

The analysis of the interference pattern is simple if the line connecting the holes is perpendicular to wave vector $\mathbf{k} \approx \mathbf{k}_0$ – see Fig. 6a. Selecting the coordinate axes as shown in that figure, and using for distances $R_{1,2}$ the same expansion as in Eq. (73), for the interference term in Eq. (88) we get

$$\cos[k(R_1 - R_2) + \varphi] \approx \cos\left(\frac{kxa}{z} + \varphi\right). \quad (8.90)$$

This means that the intensity does not depend on y , i.e. the interference pattern in the plane of constant z presents straight, parallel strips, perpendicular to vector \mathbf{a} , with the period given by Eq. (60), i.e. by the Bragg law.³¹ Note that this (somewhat counter-intuitive) result is strictly valid only at $(x^2 + y^2) \ll z^2$; it is straightforward to use the next term in the Taylor expansion (73) to show that farther from the interference pattern center the strips start to diverge.

³¹ The phase shift φ vanishes at the normal incidence of a plane wave on the holes. Note, however, that the spatial shift of the interference pattern following from Eq. (90), $\Delta x = -(z/ka)\varphi$, is extremely convenient for the experimental measurement of the phase shift between two waves, especially if it is induced by some factor (such as insertion of a transparent object into one of interferometer's arms, etc.) that may be turned on/off at will.

8.6. Diffraction on a slit

Now let us use the Huygens principle to analyze a more complex problem: plane wave's diffraction on a long straight slit of constant width a (Fig. 11).

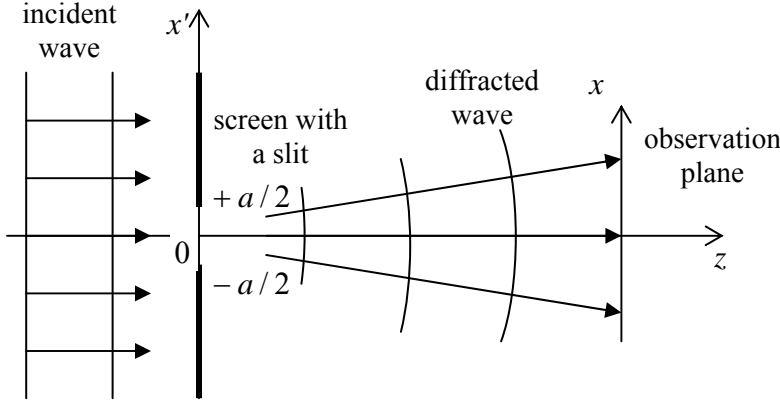


Fig. 8.11. Diffraction on a slit.

According to Eq. (70), in order to use the Huygens principle for the problem analysis we need to have $\lambda \ll a \ll z$. Moreover, the simple formulation (78) of the principle is only valid for small observation angles, $|x| \ll z$. Note, however, that the relation between two small dimensionless numbers, z/a and a/λ is so far arbitrary; as we will see in a minute, this relation will determine the type of the observed diffraction pattern.

Let us apply Eq. (78) to our current problem (Fig. 11), for the sake of simplicity assuming the normal wave incidence, and taking $z = 0$ at the screen plane:

$$f_{\omega}(x, z) = f_0 \frac{k}{2\pi i} \int_{-a}^{+a} dx' \int_{-\infty}^{+\infty} dy' \frac{\exp\left\{ik[(x-x')^2 + y'^2 + z^2]^{1/2}\right\}}{[(x-x')^2 + y'^2 + z^2]^{1/2}}, \quad (8.91)$$

where $f_0 \equiv f_{\omega}(x', 0) = \text{const}$ is the incident wave's amplitude. This is the same integral as in Eq. (72), except for the finite limits for x' , and may be simplified similarly, using the small-angle condition $(x-x')^2 + y'^2 \ll z^2$:

$$f(x, z) \approx f_0 \frac{k}{2\pi i} \frac{e^{ikz}}{z} \int_{-a/2}^{+a/2} dx' \int_{-\infty}^{+\infty} dy' \exp \frac{ik[(x-x')^2 + y'^2]}{2z} = f_0 \frac{k}{2\pi i} \frac{e^{ikz}}{z} I_x I_y. \quad (8.92)$$

The integral over y is the same as in the last section:

$$I_y \equiv \int_{-\infty}^{+\infty} \exp \frac{iky'^2}{2z} dy' = \left(\frac{2\pi iz}{k} \right)^{1/2} \quad (8.93)$$

but the integral over x is more complicated, because of its finite limits:

$$I_x \equiv \int_{-a/2}^{+a/2} \exp \frac{ik(x-x')^2}{2z} dx'. \quad (8.94)$$

It may be simplified in the following two (opposite) limits.

(i) *Fraunhofer diffraction* takes place when $z/a \gg a/\lambda$ - the relation which may be rewritten either as $a \ll (z\lambda)^{1/2}$, or as $ka^2 \ll z$. In this limit the ratio kx'^2/z is negligibly small for all values of x' under the integral, and we can approximate it as

$$\begin{aligned} I_x &= \int_{-a/2}^{+a/2} \exp \frac{ik(x^2 - 2xx' + x'^2)}{2z} dx' \approx \int_{-a/2}^{+a/2} \exp \frac{ik(x^2 - 2xx')}{2z} dx' \\ &= \exp \frac{ikx^2}{2z} \int_{-a/2}^{+a/2} \exp \left\{ -\frac{ikxx'}{z} \right\} dx' = \frac{2z}{kx} \exp \left\{ \frac{ikx^2}{2z} \right\} \sin \frac{kxa}{2z}, \end{aligned} \quad (8.95)$$

so that Eq. (92) yields

$$f_\omega(x, z) \approx f_0 \frac{k}{2\pi i} \frac{e^{ikz}}{z} \frac{2z}{kx} \left(\frac{2\pi i z}{k} \right)^{1/2} \exp \left\{ \frac{ikx^2}{2z} \right\} \sin \frac{kxa}{2z}, \quad (8.96)$$

and hence the relative wave intensity is

Fraunhofer
diffraction
pattern

$$\frac{\bar{S}(x, z)}{S_0} = \left| \frac{f_\omega(x, z)}{f_0} \right|^2 = \frac{8z}{\pi k x^2} \sin^2 \frac{kxa}{2z} = \frac{2}{\pi} \frac{ka^2}{z} \text{sinc}^2 \left(\frac{ka\alpha}{2} \right), \quad (8.97)$$

where S_0 is the (average) intensity of the incident wave, and $\alpha \equiv x/z \ll 1$ is the scattering angle. Comparing this expression with Eq. (69), we see that this the diffraction pattern is exactly the same as that of a similar (uniform, 1D) object in the Born approximation – see the red line in Fig. 8. Note again that the angular width $\delta\alpha$ of the Fraunhofer pattern is of the order of $1/ka$, so that its linear width $\delta x = z\delta\alpha \sim z/ka \sim z\lambda/a$.³² Hence the condition of the Fraunhofer approximation validity may be also presented as $a \ll \delta x$.

(ii) *Fresnel diffraction*. In the opposite limit of a relatively wide slit, with $a \gg \delta x = z\delta\alpha \sim z/ka \sim z\lambda/a$, i.e. $ka^2 \gg z$, the diffraction patterns at two slit edges are well separated. Hence, near each edge (for example, near $x' = -a/2$) we may simplify Eq. (94) as

$$I_x(x) \approx \int_{-a/2}^{+\infty} \exp \frac{ik(x-x')^2}{2z} dx' = \left(\frac{2z}{k} \right)^{1/2} \int_{(k/2z)^{1/2}(x+a/2)}^{+\infty} \exp \{ i\zeta^2 \} d\zeta, \quad (8.98)$$

and express it via the special functions called the *Fresnel integrals*:³³

Fresnel
integrals

$$\mathcal{C}(\xi) \equiv \left(\frac{2}{\pi} \right)^{1/2} \int_0^\xi \cos(\zeta^2) d\zeta, \quad \mathcal{S}(\xi) \equiv \left(\frac{2}{\pi} \right)^{1/2} \int_0^\xi \sin(\zeta^2) d\zeta, \quad (8.99)$$

whose plots are shown in Fig. 12. As was mentioned above, at large values of their argument (ξ), both functions tend to $1/2$.

³² Note also that since in this limit $ka^2 \ll z$, Eq. (97) shows that even the maximum value $S(0, z)$ of the diffracted wave intensity is much less than intensity S_0 of the incident wave. This is natural, because the incident power $S_0 a$ per unit length of the slit is now distributed over a much larger width $\delta x \gg a$, so that $S(0, z) \sim S_0 (a/\delta x) \ll S_0$.

³³ Slightly different definitions of these functions, mostly affecting constant factors, may also be met in literature.

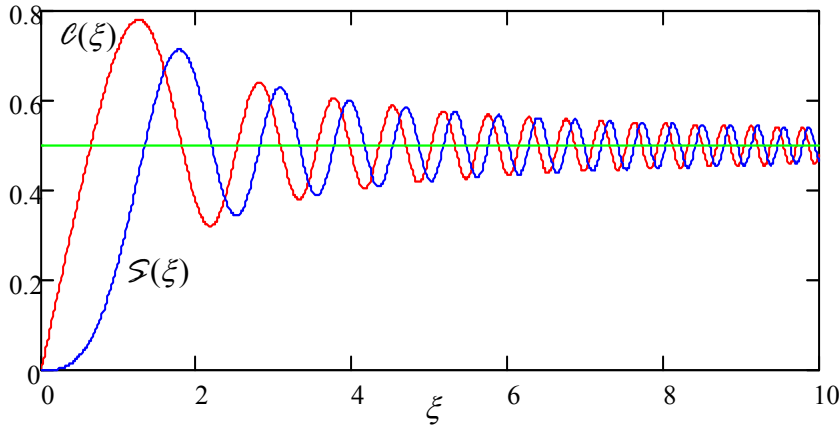


Fig. 8.12. Fresnel integrals.

Plugging this expression into Eq. (92) and (98), for the diffracted wave intensity, in the Fresnel limit (i.e. at $|x + a/2| \ll a$), we get

$$\frac{\bar{S}(x, z)}{S_0} = \frac{1}{2} \left\{ \left[\mathcal{C}\left(\left(\frac{k}{2z}\right)^{1/2} \left(x + \frac{a}{2}\right)\right) + \frac{1}{2} \right]^2 + \left[\mathcal{S}\left(\left(\frac{k}{2z}\right)^{1/2} \left(x + \frac{a}{2}\right)\right) + \frac{1}{2} \right]^2 \right\}. \quad (8.100)$$

Fresnel diffraction pattern

A plot of this function (Fig. 13) shows that the diffraction pattern is very peculiar: while in the “shade” region $x < -a/2$ the wave intensity fades monotonically, the transition to the “light” region within the gap ($x > -a/2$) is accompanied by intensity oscillations, just as at the Fraunhofer diffraction – cf. Fig. 8.

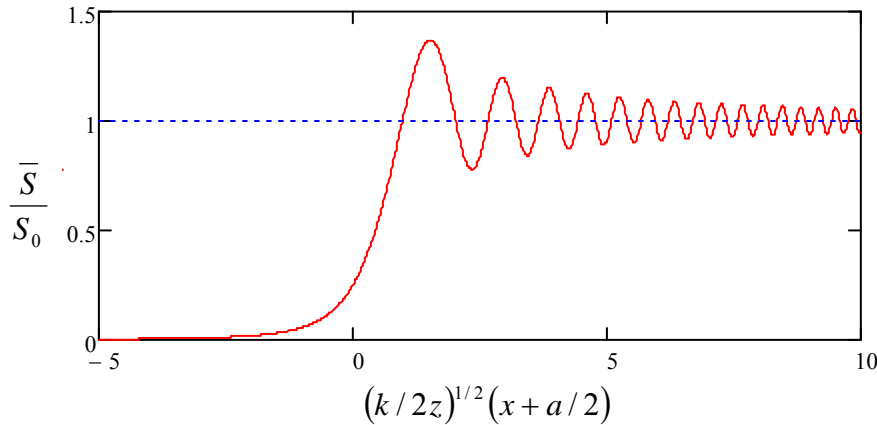


Fig. 8.13. Fresnel diffraction pattern.

This behavior, which is described by the following asymptotes,

$$\frac{\bar{S}}{S_0} \rightarrow \begin{cases} 1 + \frac{1}{\sqrt{\pi}} \frac{\sin(\xi^2 - \pi/4)}{\xi}, & \text{for } \xi \equiv \left(\frac{k}{2z}\right)^{1/2} \rightarrow +\infty, \\ \frac{1}{4\pi\xi^2}, & \text{for } \xi \rightarrow -\infty, \end{cases} \quad (8.101)$$

is essentially an artifact of observing just the wave intensity (i.e. its real amplitude) rather than its phase as well. Indeed, as may be seen even more clearly from the parametric presentation of the Fresnel

integrals (Fig. 14), these functions oscillate similarly at large positive and negative values of their argument. Physically, this means that the wave diffraction by the slit edge leads to similar oscillations of its phase at $x > -a/2$ and $x < -a/2$; however, in the latter region (i.e. inside the slit) the diffracted wave overlaps the incident wave passing through the slit directly, and their interference reveals the phase oscillations, making them visible in the measured intensity as well.

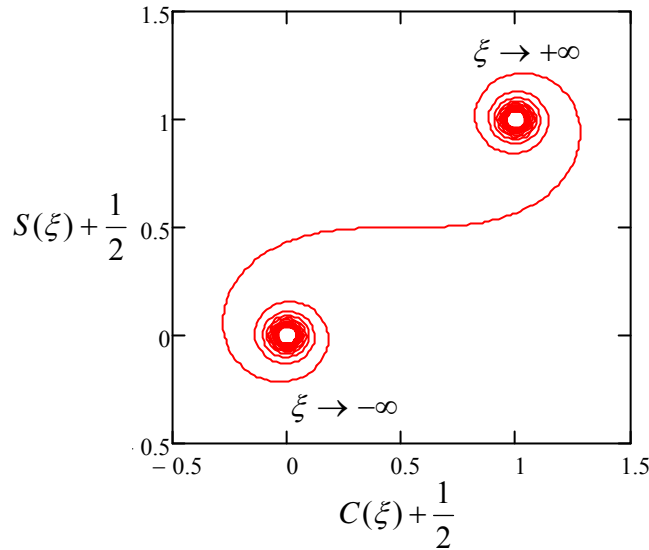


Fig. 8.14. Parametric representation of the Fresnel integrals. This pattern is called either the *Euler spiral* or *Cornu spiral*.

Note that according to Eq. (100), the linear scale of the Fresnel diffraction pattern is $(2z/k)^{1/2}$, i.e. is complied with estimate (77). If the slit is gradually narrowed, so that width a becomes comparable to that scale,³⁴ the Fresnel interference patterns from both edges start to “collide” (interfere). The resulting wave, fully described by Eq. (94), is just a sum of two contributions of the type (98) from the both edges of the slit. The resulting interference pattern is somewhat complicated, and only $a \ll \delta x$ it is reduced to the simple Fraunhofer pattern (97). Of course, this crossover from the Fresnel to Fraunhofer diffraction may be also observed, at fixed wavelength λ and slit width a , by increasing z , i.e. by measuring the diffraction pattern farther and farther from the slit.

Note that the Fraunhofer limit is always valid if the diffraction measured as a function of the diffraction angle α alone, i.e. effectively at infinity, $z \rightarrow \infty$. This may be done, for example, by collecting the diffracted wave with a “positive” (converging) lense, and observing the diffraction pattern in its focal plane.

8.7. Geometrical optics placeholder

Behind all these details, I would not like the reader to miss the main feature of diffraction, that has an overwhelming practical significance. Namely, besides narrow diffraction “cones” (actually, parabolic-shaped regions) with lateral scale $\Delta x \sim (\lambda z)^{1/2}$, the wave far behind a slit of width $a \gg \lambda$ repeats the field just behind the slit, i.e. reproduces the unperturbed incident wave inside the slit, and has negligible intensity in the shade regions outside it. An evident generalization of this fact is that when a plane wave (in particular an electromagnetic wave) passes any opaque object of large size $a \gg \lambda$, it propagates around it, by distances z up to $\sim (a/\lambda)^{1/2}$, along straight lines, with virtually negligible

³⁴ Note that this condition may be also rewritten as $a \sim \delta x$, i.e. $z/a \sim a/\lambda$.

diffraction effects. This fact gives the strict foundation for the very notion of the wave *ray* (or *beam*), as the line perpendicular to the local front of a quasi-plane wave. In a uniform media such ray is a straight line, but changes in accordance with the Snell law at the interface of two media with different wave speed v , i.e. different values of the refraction index. The notion of rays enables the whole field of geometric optics, devoted mostly to ray tracing in various (sometimes very complex) systems.

This is why, at this point, an E&M course that followed the scientific logic more faithfully than this one, would give an extended discussion of the geometric and quasi-geometric optics, including (as a minimum³⁵) such vital topics as

- the so-called *lensmaker's equation* expressing the focus length f of a lens via the curvature radii of its spherical surfaces and the refraction index of the lens material,
- the *thin lense formula* relating the image distance from the lense via f and the source distance,
- the concepts of basic optical instruments such as *telescopes* and *microscopes*,
- the concepts of the spherical, angular, and chromatic *aberrations* (image distortions);
- wave effects in optical instruments, including the so-called *Abbe limit*³⁶ on the focal spot size.³⁷

However, since I have made a (possibly, wrong) decision to follow the common tradition in selecting the main topics for this course, I do not have time left for such discussion. Still, I am placing this “placeholder” pseudo-section to relay my conviction that any educated physicist has to know the geometric optics basics. If the reader has not had an exposure to this subject during his or her undergraduate studies, I highly recommend at least browsing one of available textbooks.³⁸

8.8. Fraunhofer diffraction from more complex scatterers

So far, our discussion of diffraction has been limited to a very simple geometry – a single slit in an otherwise opaque screen (Fig. 11). However, in the most important Fraunhofer limit, $z \gg ka^2$, it is easy to get a very simple expression for the plane wave diffraction/interference by a plane orifice (with linear size $\sim a$) of an arbitrary shape. Indeed, the evident 2D generalization of approximation (93)-(94) is

$$I_x I_y = \int_{\text{orifice}} \exp \frac{ik[(x-x')^2 + (y-y')^2]}{2z} dx' dy' \quad (8.102)$$

$$\approx \exp \left\{ \frac{ik(x^2 + y^2)}{2z} \right\} \int_{\text{orifice}} \exp \left\{ -i \frac{kxx'}{z} - i \frac{kyy'}{z} \right\} dx' dy',$$

³⁵ Admittedly, even this list leaves aside several spectacular effects due to crystal anisotropy, including such a beauty as *conical refraction* in biaxial crystals - see, e.g., Chapter 15 of the classical textbook by M. Born and E. Wolf, cited in the end of Sec. 7.1.

³⁶ Reportedly, due to not only E. Abbe (1873), but also to H. von Helmholtz (1874).

³⁷ In contrast to other topics of this list, whose study may be based on the ray approach, i.e. on purely geometric optics, the description of these effects requires at least an approximate account of wave properties of light. Such account may be based either on the Huygens principle or on the so-called *paraxial equation*

$$\partial a / \partial z = (1/2ik) \nabla_{x,y}^2 a,$$

for the complex amplitude $a(\mathbf{r})$ of the field represented in the form $f(\mathbf{r}) = a(\mathbf{r})e^{ikz}$. The paraxial approximation follows from the Helmholtz equation (7.192) in essentially the same limit ($|\nabla a| \ll k$; $|x|, |y| \ll z$) as Eq. (78).

³⁸ My top recommendation for that purpose would be Chapters 3-6 and Sec. 8.6 in Born and Wolf. A simpler alternative is Chapter 10 in G. R. Fowles, *Introduction to Modern Optics*, 2nd ed., Dover, 1989.

so that besides the inconsequential total phase factor, Eq. (92) is reduced to

$$f(\mathbf{p}) \propto f_0 \int_{\text{orifice}} \exp\{-i\mathbf{\kappa} \cdot \mathbf{p}'\} d^2 \rho' = f_0 \int_{\text{screen}} T(\mathbf{p}') \exp\{-i\mathbf{\kappa} \cdot \mathbf{p}'\} d^2 \rho', \quad (8.103)$$

where the 2D vector $\mathbf{\kappa}$ (not to be confused with wave vector \mathbf{k} that is virtually perpendicular to $\mathbf{\kappa}$!) is defined as

$$\mathbf{\kappa} \equiv k \frac{\mathbf{p}}{z} \approx \mathbf{q} \equiv \mathbf{k} - \mathbf{k}_0, \quad (8.104)$$

$\mathbf{p} = \{x, y\}$ and $\mathbf{p}' = \{x', y'\}$ are 2D radius-vectors in, respectively, the observation and screen planes (both nearly normal to vectors \mathbf{k} and \mathbf{k}_0), function $T(\mathbf{p}')$ describes screen's transparency at point \mathbf{p}' , and the last integral in Eq. (103) is over the whole screen plane $z' = 0$. (Though the strict equivalence of the two forms of Eq. (103) is only valid if $T(\mathbf{p}')$ equals to either 1 or 0, its last form may be readily obtained from Eq. (78) with $f(\mathbf{r}') = T(\mathbf{p}') f_0$ for any transparency profile, provided that $T(\mathbf{p}')$ is an arbitrary function but changes only at distances much larger than $\lambda \equiv 2\pi/k$.)

From the mathematical point of view, the last form of Eq. (103) is the 2D spatial Fourier transform of function $T(\mathbf{p}')$, with the reciprocal variable $\mathbf{\kappa}$ revealed by the observation point position: $\mathbf{p} = (z/k)\mathbf{\kappa} = (z\lambda/2\pi)\mathbf{\kappa}$. This interpretation is useful because of the experience we all have with the Fourier transform, mostly in the context of its time/frequency applications. For example, if the orifice is a single small hole, $T(\mathbf{p}')$ may be approximated by a delta-function, so that Eq. (103) yields $f(\mathbf{p}) \approx \text{const}$. This corresponds (at least for the small diffraction angles $\alpha \equiv \rho/z$, for which the Huygens approximation is valid) to a spherical wave spreading from the point-like orifice. Next, for two small holes, Eq. (103) immediately gives the Young interference pattern (90). Let me now use Eq. (103) to analyze the simplest (and most important) 1D transparency profiles, leaving 2D cases for reader's exercise.

(i) A single slit of width a (Fig. 11) may be described by transparency

$$T(\mathbf{p}') = \begin{cases} 1, & \text{for } |x'| < a/2, \\ 0, & \text{otherwise.} \end{cases} \quad (8.105)$$

Its substitution into Eq. (103) yields

$$f(\mathbf{p}) \propto f_0 \int_{-a/2}^{+a/2} \exp\{-i\kappa_x x'\} dx' = f_0 \frac{\exp\{-i\kappa_x a/2\} - \exp\{i\kappa_x a/2\}}{-i\kappa_x} \propto \text{sinc}\left(\frac{\kappa_x a}{2}\right) = \text{sinc}\left(\frac{kx a}{2z}\right), \quad (8.106)$$

naturally returning us to Eqs. (64) and (97), and hence to the red lines in Fig. 8 for the wave intensity. (Please note again that Eq. (103) describes only the Fraunhofer, but not the Fresnel diffraction!)

(ii) Two narrow similar, parallel slits with a much larger distance a between them, may be described by taking

$$T(\mathbf{p}') \propto \delta(x' - a/2) + \delta(x' + a/2), \quad (8.107)$$

so that Eq. (103) yields the generic interference pattern,

$$f(\mathbf{p}) \propto f_0 \left[\exp\left\{-\frac{i\kappa_x a}{2}\right\} + \exp\left\{\frac{i\kappa_x a}{2}\right\} \right] \propto \cos \frac{\kappa_x a}{2} = \cos \frac{kx a}{2z}, \quad (8.108)$$

whose intensity is shown with the blue line in Fig. 8.

(iii) In a more realistic Young-type two-slit experiment, each slit has width (say, w) which is much larger than light wavelength λ , but still much smaller than slit spacing a . This situation may be described by the following transparency function

$$T(\mathbf{p}') = \sum_{\pm} \begin{cases} 1, & \text{for } |x' \pm a/2| < w/2, \\ 0, & \text{otherwise,} \end{cases} \quad (8.109)$$

for which Eq. (103) yields a natural combination of results (106) (with a replaced with w) and (108):

$$f(\mathbf{r}) \propto \text{sinc}\left(\frac{kxw}{2z}\right) \cos\left(\frac{kxa}{2z}\right). \quad (8.110)$$

This is the usual interference pattern modulated by a Fraunhofer-diffraction envelope (shown with the dashed blue line Fig. 15). Since function $\text{sinc}^2 \xi$ decreases very fast beyond its first zeros at $\xi = \pm\pi$, the practical number of observable interference fringes is close to $2a/w$.

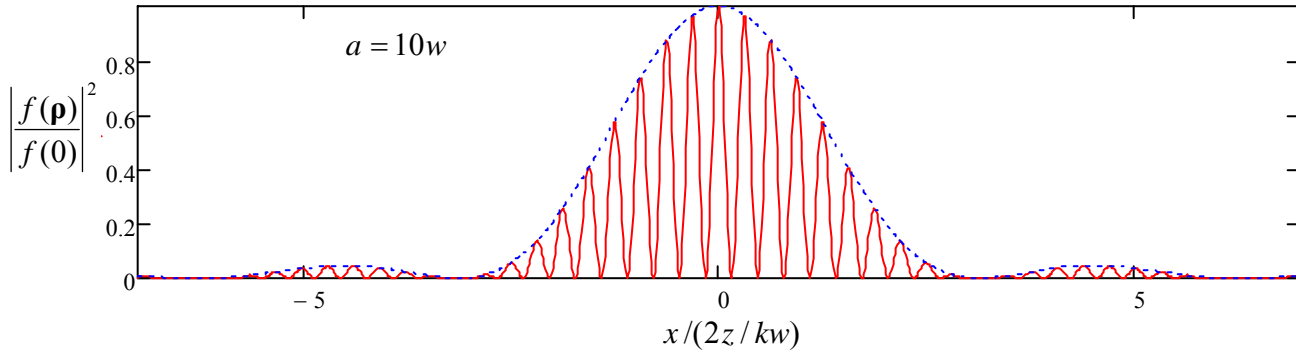


Fig. 8.15. Young's double-slit interference pattern for a finite slit width.

(iv) A structure very useful for experimental and engineering practice is a set of many parallel slits, called the *diffraction grating*.³⁹ Indeed, if the slit width is much less than the grating period d , then the transparency function may be approximated as

$$T(\mathbf{p}') \propto \sum_{n=-\infty}^{+\infty} \delta(x' - nd), \quad (8.111)$$

and Eq. (103) yields

$$f(\mathbf{p}) \propto \sum_{n=-\infty}^{n=+\infty} \exp\{-in\kappa_x d\} = \sum_{n=-\infty}^{n=+\infty} \exp\left\{-i \frac{nkxd}{z}\right\}. \quad (8.112)$$

This sum vanishes for all values of $\kappa_x d$ that are not multiples of 2π , so that the result describes sharp intensity peaks at diffraction angles

³⁹ The rudimentary diffraction grating effect, produced by parallel fibers of bird feathers, was discovered as early as in 1673 by J. Gregory - who has also invented the reflecting ("Gregorian") telescope.

$$\alpha_m \equiv \left(\frac{x}{z} \right)_m = \left(\frac{\kappa_x}{k} \right)_m = \frac{2\pi}{kd} m = \frac{\lambda}{d} m. \quad (8.113)$$

Taking into account that this result is only valid for small angles $|\alpha_m| \ll 1$, it may be interpreted exactly as Eq. (59) – see Fig. 6a. However, in contrast with the interference (108) from two slits, the destructive interference from many slits kills the net wave as soon as the angle is even slightly different from each Bragg angle (60). This is very convenient for spectroscopic purposes, because the diffraction lines produced by multi-frequency waves do not overlap even if the frequencies of their adjacent components are very close.

Two features of practical diffraction gratings make their properties different from this simple picture. First, the finite number N of slits, which may be described by limiting sum (109) to interval $n = [-N/2, +N/2]$, results in the finite spread, $\delta\alpha/\alpha \sim 1/N$, of each diffraction peak, and hence in the reduction of grating's spectral resolution. (Unintentional variations of the inter-slit distance d have a similar effect, so that before the advent of high-resolution photolithography, special high-precision mechanical tools have been used for grating fabrication.)

Second, the finite slit width w leads to the diffraction peak pattern modulation by a $\text{sinc}^2(kw\alpha/2)$ envelope, similarly to pattern shown in Fig. 15. Actually, for spectroscopic purposes such modulation is a plus, because only one diffraction peak (say, with $m = \pm 1$) is practically used, and if the frequency spectrum of the analyzed wave is very broad (cover more than one octave), the higher peaks produce undesirable hindrance. Because of this reason, w is frequently selected to be equal exactly to $d/2$, thus suppressing each other diffraction maximum. Moreover, sometimes semi-transparent films are used to make the transparency function $T(\mathbf{r}')$ continuous and close to the sinusoidal one:

$$T(\mathbf{p}') \approx T_0 + T_1 \cos \frac{2\pi x'}{d} = T_0 + \frac{T_1}{2} \left(\exp \left\{ i \frac{2\pi x'}{d} \right\} + \exp \left\{ -i \frac{2\pi x'}{d} \right\} \right). \quad (8.114)$$

Plugging the last expression into Eq. (103) and integrating, we see that the output wave consists of just 3 components: the direct-passing wave (proportional to T_0) and two diffracted waves (proportional to T_1) propagating in the directions of the two lowest Bragg angles, $\alpha_{\pm 1} = \pm \lambda/d$.

Relation (103) may be also readily used to obtain one more general (and rather curious) result called the *Babinet principle*. Consider two experiments with diffraction of similar plane waves on two “complementary” screens who together would cover the whole plane, without a hole or an overlap. (Think, for example, about an opaque disk of radius R and a large opaque screen with a round orifice of the same radius.) Then, according to the Babinet principle, the diffracted wave patterns produced by these two screens in all directions with $\alpha \neq 0$ are *identical*. The proof of this principle is straightforward: since the transparency functions produced by the screens are complementary in the following sense:

$$T(\mathbf{p}') \equiv T_1(\mathbf{p}') + T_2(\mathbf{p}') = 1, \quad (8.115)$$

and (in the Fraunhofer approximation (103) only!) the diffracted wave is a linear Fourier transform of $T(\mathbf{p}')$, we get

$$f_1(\mathbf{p}) + f_2(\mathbf{p}) = f_0(\mathbf{p}), \quad (8.116)$$

where f_0 is the wave “scattered” by the composite screen with $T_0(\mathbf{p}') \equiv 1$, i.e. the unperturbed initial wave propagating in the initial direction ($\alpha = 0$). In all other directions, $f_1 = -f_2$, i.e. the diffracted waves

are indeed similar besides the difference in sign - which is equivalent to a phase shift by $\pm\pi$. However, it is important to remember that the Babinet principle notwithstanding, in real experiments the diffracted waves may interfere with the unperturbed plane wave $f_0(\mathbf{p})$, leading to different diffraction pattern in cases 1 and 2 – see, e.g., Fig. 13 and its discussion.

8.9. Magnetic dipole and electric quadrupole radiation

Throughout this chapter, we have seen how many important results may be obtained from Eq. (26) for the electric dipole radiation by a small-size source (Fig. 1). Only in rare cases when such radiation is absent, for example if the dipole moment \mathbf{p} of the source equals zero (or does not change at time – either at all, or at the frequency of our interest), higher-order effects may be important. I will discuss the main two of them, the quadrupole electric and dipole magnetic radiation – mostly for reference purposes, because we would not have much time to discuss their applications.

In Sec. 2 above, the electric dipole radiation was calculated by plugging the first, leading term of expansion (19) into the exact formula (17b) for the retarded vector-potential $\mathbf{A}(\mathbf{r}, t)$. Let us make a more exact calculation, by keeping the second term of that expansion as well:

$$\mathbf{j}\left(\mathbf{r}', t - \frac{R}{v}\right) \approx \mathbf{j}\left(\mathbf{r}', t - \frac{r}{v} + \frac{\mathbf{r}' \cdot \mathbf{n}}{v}\right) = \mathbf{j}\left(\mathbf{r}', t' + \frac{\mathbf{r}' \cdot \mathbf{n}}{v}\right), \quad \text{where } t' \equiv t - \frac{r}{v}. \quad (8.117)$$

Since the expansion is only valid if the last term in the second argument is relatively small, in the Taylor expansion of \mathbf{j} with respect to that argument we may keep just the first two leading terms:

$$\mathbf{j}\left(\mathbf{r}', t - \frac{R}{v}\right) \approx \mathbf{j}(\mathbf{r}', t') + \frac{1}{v} \frac{\partial}{\partial t'} \mathbf{j}(\mathbf{r}', t') (\mathbf{r}' \cdot \mathbf{n}), \quad (8.118)$$

so that Eq. (17b) yields $\mathbf{A} = \mathbf{A}_e + \mathbf{A}'$, where \mathbf{A}_e is the electric dipole contribution as given by Eq. (23), and \mathbf{A}' is the new term of the next order in small parameter $r' \ll r$:

$$\mathbf{A}'(\mathbf{r}, t) = \frac{\mu}{4\pi r v} \frac{\partial}{\partial t'} \int \mathbf{j}(\mathbf{r}', t') (\mathbf{r}' \cdot \mathbf{n}) d^3 r'. \quad (8.119)$$

Just as was done in Sec. 2, let us evaluate this term for a system of nonrelativistic particles with electric charges q_k and radius-vectors $\mathbf{r}_k(t)$:

$$\mathbf{A}'(\mathbf{r}, t) = \frac{\mu}{4\pi r v} \left[\frac{d}{dt} \sum_k q_k \dot{\mathbf{r}}_k (\mathbf{r}_k \cdot \mathbf{n}) \right]_{t=t'}. \quad (8.120)$$

Using the “bac minus cab” identity of the vector algebra again,⁴⁰ Eq. (120) may be rewritten as

$$\begin{aligned} \dot{\mathbf{r}}_k (\mathbf{r}_k \cdot \mathbf{n}) &= \frac{1}{2} \dot{\mathbf{r}}_k (\mathbf{r}_k \cdot \mathbf{n}) + \frac{1}{2} \dot{\mathbf{r}}_k (\mathbf{n} \cdot \mathbf{r}_k) = \frac{1}{2} (\mathbf{r}_k \times \dot{\mathbf{r}}_k) \times \mathbf{n} + \frac{1}{2} \mathbf{r}_k (\mathbf{n} \cdot \dot{\mathbf{r}}_k) + \frac{1}{2} \dot{\mathbf{r}}_k (\mathbf{n} \cdot \mathbf{r}_k) \\ &= \frac{1}{2} (\mathbf{r}_k \times \dot{\mathbf{r}}_k) \times \mathbf{n} + \frac{1}{2} \frac{d}{dt} [\mathbf{r}_k (\mathbf{n} \cdot \mathbf{r}_k)], \end{aligned} \quad (8.121)$$

so that the right-hand part of Eq. (120) may be presented as a sum of two terms, $\mathbf{A}' = \mathbf{A}_m + \mathbf{A}_q$, where

⁴⁰ If you need, see, e.g., MA Eq. (7.5).

$$\mathbf{A}_m(\mathbf{r}, t) = \frac{\mu}{4\pi r v} \dot{\mathbf{m}}(t') \times \mathbf{n} = \frac{\mu}{4\pi r v} \dot{\mathbf{m}}\left(t - \frac{r}{v}\right) \times \mathbf{n}, \quad \text{with } \mathbf{m}(t) \equiv \frac{1}{2} \sum_k \mathbf{r}_k(t) \times q_k \dot{\mathbf{r}}_k(t), \quad (8.122)$$

$$\mathbf{A}_q(\mathbf{r}, t) = \frac{\mu}{8\pi r v} \left[\frac{d^2}{dt^2} \sum_k q_k \mathbf{r}_k (\mathbf{n} \cdot \mathbf{r}_k) \right]_{t=t'}. \quad (8.123)$$

Comparing the second of Eqs. (122) with Eq. (5.91), we see that \mathbf{m} is just the magnetic moment of the source. On the other hand, the first of Eqs. (122) is absolutely similar in structure to Eq. (23), with \mathbf{p} replaced by $(\mathbf{m} \times \mathbf{n})/v$, so that for the corresponding component of the magnetic field it gives (in the same approximation $r \gg \lambda$) the result similar to Eq. (24):

Magnetic
dipole
radiation
field

$$\mathbf{B}_m(\mathbf{r}, t) = \frac{\mu}{4\pi r v} \nabla \times \left[\dot{\mathbf{m}}\left(t - \frac{r}{v}\right) \times \mathbf{n} \right] = -\frac{\mu}{4\pi r v^2} \mathbf{n} \times \left[\ddot{\mathbf{m}}\left(t - \frac{r}{v}\right) \times \mathbf{n} \right]. \quad (8.124)$$

According to this expression, just as at the electric dipole radiation, vector \mathbf{B} is perpendicular to vector \mathbf{n}_r , and its magnitude is also proportional to the $\sin\theta$, where θ is now the angle between the direction toward the observation point and the second time derivative of vector \mathbf{m} rather than \mathbf{p} :

$$B_m = \frac{\mu}{4\pi r v^2} \ddot{m}\left(t - \frac{r}{v}\right) \sin\theta. \quad (8.125)$$

As the result, the intensity of this *magnetic dipole radiation* has the similar angular distribution:

Magnetic
dipole
radiation
power
density

$$S_r = ZH^2 = \frac{Z}{(4\pi v^2 r)^2} \left[\ddot{m}\left(t - \frac{r}{v}\right) \right]^2 \sin^2\theta \quad (8.126)$$

- cf. Eq. (26). Note, however, that this radiation is usually much weaker than its electric counterpart. For example, for a nonrelativistic particle with electric charge q , moving on a trajectory with of size $\sim a$, the electric dipole moment is of the order of qa , while its magnetic moment scales as $qa^2\omega$, where ω is the motion frequency. As a result, the ratio of the magnetic and electric dipole radiation intensities is of the order of $(a\omega/v)^2$, i.e. the squared ratio of particle's speed to the speed of emitted waves – that has to be much smaller than 1 for our nonrelativistic estimate to be valid.

The angular distribution of the *electric quadrupole radiation*, described by Eq. (123), is more complicated. In order to show this, we may add to \mathbf{A}_q a vector parallel to \mathbf{n} (i.e. along the wave propagation), getting

$$\mathbf{A}_q(\mathbf{r}, t) \rightarrow \frac{\mu}{24\pi r v} \ddot{\mathbf{Q}}\left(t - \frac{r}{v}\right), \quad \text{where } \mathbf{Q} \equiv \sum_k q_k \{3\mathbf{r}_k(\mathbf{n} \cdot \mathbf{r}_k) - \mathbf{n}r_k^2\}, \quad (8.127)$$

because this addition does not give any contribution to the transverse component of the electric and magnetic fields, i.e. to the radiated wave. According to the above definition of vector \mathbf{Q} , its Cartesian components may be presented as⁴¹

⁴¹ In electrostatics, the symmetric, zero-trace tensor Q determines the next term in the potential expansion (3.5):

$$\phi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \left(\frac{1}{r} Q + \frac{1}{r^2} \sum_{j=1}^3 n_j p_j + \frac{1}{2r^3} \sum_{j,j'=1}^3 n_j n_{j'} Q_{jj'} + \dots \right).$$

$$Q_j = \sum_{j'=1}^3 Q_{jj'} n_{j'}, \quad (8.128)$$

where $Q_{jj'}$ are elements of the so-called *electric quadrupole tensor* Q of the system:⁴²

$$Q_{jj'} = \sum_k q_k (3r_j r_{j'} - r^2 \delta_{jj'})_k. \quad (8.129a)$$

For clarity, let me spell out the tensor in its matrix form:

$$Q = \sum_k q_k \begin{pmatrix} 2x^2 - y^2 - z^2 & 3xy & 3xz \\ 3xy & 2y^2 - x^2 - z^2 & 3yz \\ 3xz & 3yz & 2z^2 - x^2 - y^2 \end{pmatrix}_k. \quad (8.129b)$$

Differentiating the first of Eqs. (127) at $r \gg \lambda$, we get

$$\mathbf{B}_q(\mathbf{r}, t) = -\frac{\mu}{24\pi r v^2} \mathbf{n} \times \ddot{\mathbf{Q}}\left(t - \frac{r}{v}\right). \quad (8.130)$$

Electric
quadrupole
radiation
field

Superficially, this expression is similar to Eqs. (24) or (124), but according to Eqs. (127) and (129), components of vector \mathbf{Q} depend on the direction of vector \mathbf{n} , leading to a different angular dependence of S_r .

As the simplest example, let us consider a system of two equal point electric charges moving at equal distances $d(t) \ll \lambda$ from a stationary center (Fig. 16).

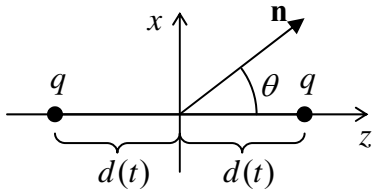


Fig. 8.16. The simplest system emitting electric quadrupole radiation.

Due to the symmetry of the system, its dipole moments \mathbf{p} and \mathbf{m} (and hence its electric and magnetic dipole radiation) vanish, but the quadrupole tensor (129) still has nonvanishing components. With the coordinate choice shown in Fig. 16, these components are diagonal:

$$Q_{xx} = Q_{yy} = -2qd^2, \quad Q_{zz} = 4qd^2. \quad (8.131)$$

With axis x in the plane of the direction \mathbf{n} toward the source (Fig. 16), so that $n_x = \sin\theta$, $n_y = 0$, $n_z = \cos\theta$, Eq. (128) yields

$$Q_x = -2qd^2 \sin\theta, \quad Q_y = 0, \quad Q_z = 4qd^2 \cos\theta, \quad (8.132)$$

⁴² As a math reminder, tensor is a matrix describing a physical reality independent of the reference frame choice, so that the Cartesian elements of the tensor have to change according to certain geometric rules if the reference frame is changed - e.g., rotated. This notion is very similar to a physical vector, that may be described by an ordered set of its Cartesian components, which change according to certain rules as the result of the reference frame' change. We may be confident that a matrix represents a tensor if it provides a linear relation between components of two physical vectors – such a \mathbf{Q} and \mathbf{n} in Eq. (128).

so that the vector product in Eq. (130) has only one nonvanishing Cartesian component:

$$(\mathbf{n} \times \ddot{\mathbf{Q}})_y = n_z \ddot{Q}_x - n_x \ddot{Q}_z = -6q \sin \theta \cos \theta \frac{d^3}{dt^3} [d^2(t)]. \quad (8.133)$$

As a result, the radiation intensity is proportional to $\sin^2 \theta \cos^2 \theta$, i.e. vanishes not only along the symmetry axis (as the dipole radiation does), but also in all directions perpendicular to this axis, reaching its maximum at $\theta = \pi/4$.

For more complex systems, the angular distribution of the electric quadrupole radiation may be different, but its total power may be always presented in a simple form

Electric
quadrupole
radiation
power

$$\mathcal{P}_q = \frac{Z}{1440\pi\nu^4} \sum_{j,j'=1}^3 (\ddot{Q}_{jj'})^2. \quad (8.134)$$

Let me finish this section by giving, without proof, one more fact important for applications: due to their different spatial structure, the magnetic dipole and electric quadrupole radiation fields do not interfere, i.e. the total power of radiation (neglecting higher multipole terms) may be found as the sum of these components, calculated independently.

8.10. Exercise problems

8.1.* In the electric dipole approximation, calculate the angular distribution and total power of electromagnetic radiation by the following classical model of the hydrogen atom: an electron rotating, at a constant distance r , about a much heavier proton. Use the latter result to evaluate the classical lifetime of the atom, borrowing the initial value of R from quantum mechanics: $R(0) = r_B \approx 0.53 \times 10^{-10}$ m.

8.2. A nonrelativistic particle of mass m with the electric charge q is placed into a uniform magnetic field \mathbf{B} . Derive the law of decrease of particle's kinetic energy due to its electromagnetic radiation at the *cyclotron frequency* $\omega_c = qB/m$. Evaluate the rate of such radiation cooling for electrons in a magnetic field of 1 T, and estimate the electron energy interval in which this result is qualitatively correct.

Hint: The cyclotron motion will be discussed in detail (for arbitrary particle velocities $v \sim c$) in Sec. 9.6 below, but I hope that the reader knows that in the nonrelativistic case ($v \ll c$) the above formula for ω_c may be readily obtained by combining the 2nd Newton law $m\mathbf{v}_\perp^2/R = q\mathbf{v}_\perp \mathbf{B}$ for the circular motion of the particle under the effect of the magnetic component of the Lorentz force (5.10), and the geometric relation $v_\perp = R\omega_c$. (Here \mathbf{v}_\perp is particle's velocity within the plane normal to vector \mathbf{B} .)

8.3. Solve the dipole antenna radiation problem discussed in Sec. 2 (see Fig. 3) for the optimal length $l = \lambda/2$, assuming⁴³ that the current distribution in each of its arms is sinusoidal:

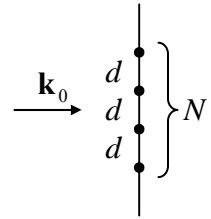
$$I(z, t) = I_0 \cos \frac{\pi z}{l} \cos \omega t.$$

⁴³ As was emphasized in Sec. 2, this is a reasonable guess rather than a controllable approximation. The exact (rather involved!) theory shows that this assumption gives errors $\sim 5\%$.

8.4. Use the harmonic oscillator model of a bound charge, given by Eq. (7.30), to explore the transition between two scattering limits discussed in Sec. 3, in particular the *resonant scattering* taking place at $\omega \approx \omega_0$. In this context, discuss the contribution of scattering into oscillator's damping.

8.5.* A sphere of radius R , made of a material with constant permanent electric polarization P_0 and mass density ρ , is free to rotate about its center of mass. Calculate the total cross-section of scattering, by the sphere, of a linearly polarized electromagnetic wave of frequency $\omega \ll R/c$, propagating in free space, in the limit of small wave amplitude, assuming that the initial orientation of the polarization vector \mathbf{P}_0 is random.

8.6. Use the Born approximation to analyze the interference pattern produced by plane wave's scattering on a set of N similar, equidistant points on a straight line normal to the direction of the incident wave's propagation – see Fig. on the right. Discuss the trend(s) of the pattern in the limit $N \rightarrow \infty$.



8.7. Use the Born approximation to calculate the differential cross-section of plane wave scattering by a dielectric cube of side a , with $\varepsilon \approx \varepsilon_0$. In the limits $ka \ll 1$ and $ka \gg 1$ (where k is the wave vector), analyze the angular dependence of the differential cross-section. Calculate the full cross-section for the simplest case when the incident wave vector is parallel to one of cube's sides.

8.8. Use the Born approximation to calculate the differential cross-section of plane wave scattering by a nonmagnetic, uniform dielectric sphere with $\varepsilon \approx \varepsilon_0$, of an arbitrary radius R . In the limits $kR \ll 1$ and $1 \ll kR$ (where k is the wave number), analyze the angular dependence of the differential cross-section, and calculate the full cross-section.

8.9. A sphere of radius R is made of a uniform, nonmagnetic, linear dielectric material. Calculate its full cross-section of scattering of a low-frequency monochromatic wave, with $k \ll 1/R$, for an arbitrary dielectric constant, and compare the result with the solution of the previous problem.

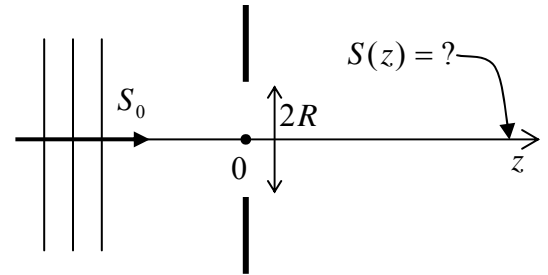
8.10. Solve the previous problem, also in the low-frequency limit $kR \ll 1$, for the case when the sphere's material has a frequency-independent Ohmic conductivity, and $\varepsilon_{\text{opt}} = \varepsilon_0$, and a relatively large skin depth ($\delta_s \gg R$), and compare the results.

8.11. Use the Born approximation to calculate the differential cross-section of plane wave scattering on a right, circular cylinder of length l and radius R , for arbitrary incidence.

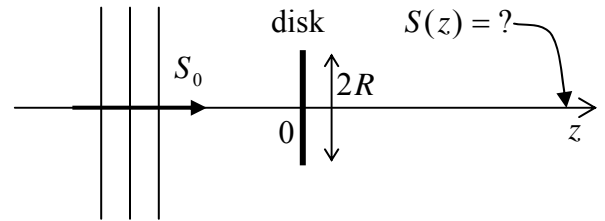
8.12. Formulate the quantitative condition of the Born approximation validity for a uniform linear-dielectric scatterer with all linear dimensions of the order of a .

8.13. Use the Huygens principle to calculate wave's intensity on the symmetry plane of the slit diffraction experiment (i.e. at $x = 0$ in Fig. 11), for arbitrary ratio z/ka^2 .

8.14. A plane wave with wavelength λ is normally incident on an opaque, plane screen, with a round orifice of radius $R \gg \lambda$. Use the Huygens principle to calculate passed wave's intensity distribution along system's symmetry axis, at distances $z \gg R$ from the screen (see Fig. on the right), and analyze the result.

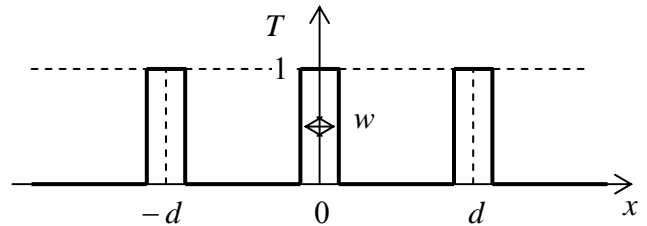


8.15. A plane monochromatic wave is normally incident on an opaque circular disk of radius $R \gg \lambda$. Use the Huygens principle to calculate wave's intensity at distance $z \gg R$ behind the disk center (see Fig. on the right). Discuss the result.



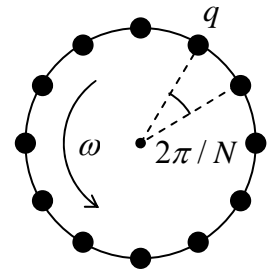
8.16. Use the Huygens principle to analyze the Fraunhofer diffraction of a plane wave normally incident on a square-shape hole, of size $a \times a$, in an opaque screen. Sketch the diffraction pattern you would observe at a sufficiently large distance, and quantify expression “sufficiently large” for this case.

8.17. Within the Fraunhofer approximation, analyze the pattern produced by a 1D diffraction grating with the periodic transparency profile shown in Fig. on the right, for the normal incidence of a plane, monochromatic wave.



8.18. N equal point charges are attached, at equal intervals, to a circle rotating with a constant angular velocity about its center – see Fig. on the right. For what values of N does the system emit:

- (i) the electric dipole radiation?
- (ii) the magnetic dipole radiation?
- (iii) the electric quadrupole radiation?



8.19. The orientation of a magnetic dipole \mathbf{m} , of a fixed magnitude, is rotating about a certain axis with angular velocity ω , with angle θ between them staying constant. Calculate the angular distribution and the average power of its radiation (into free space).

8.20. Complete the solution of the problem started in Sec. 9, by calculating the full power of radiation of the system of two charges oscillating in antiphase along the same straight line - see Fig. 6. Also, calculate the average radiation power for the case of harmonic oscillations, $d(t) = a \cos \omega t$, compare it with the case of a single charge performing similar oscillations, and interpret the difference.

This page is
intentionally left
blank

Chapter 9. Special Relativity

This chapter starts with a brief review of the special relativity's basics. This background is used, later in the chapter, for the analysis of the relation between electromagnetic field values measured in different reference frames moving relative to each other, and discussions of relativistic particle dynamics in the electric and magnetic fields, and of analytical mechanics of electromagnetism.

9.1. Einstein postulates and the Lorentz transform

As was emphasized at the derivation of expressions for the dipole and quadrupole radiation in the last chapter, they are only valid for systems of nonrelativistic particles. Thus, these results cannot be used for description of such important phenomena as the Cherenkov radiation or synchrotron radiation, in which relativistic effects are essential. Moreover, analysis of motion of charged relativistic particles in electric and magnetic fields is also a natural part of electrodynamics. This is why I will follow the tradition of using this course for a (by necessity, brief) introduction to special relativity theory. This theory is based on the idea that measurements of all physical variables (including spatial and even temporal intervals between two events) may give different results in different reference frames, in particular two frames moving relative to each other translationally (i.e. without rotation), with a certain constant velocity \mathbf{v} (Fig. 1).

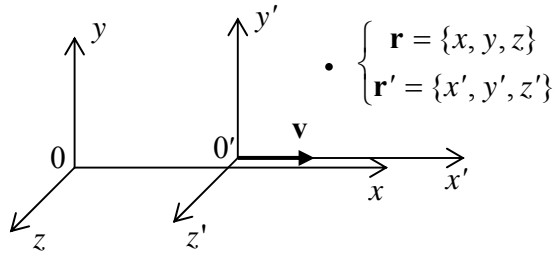


Fig. 9.1. Translational, uniform motion of two reference frames.

In the nonrelativistic (Newtonian) mechanics the problem of transfer between such reference frames has a simple solution at least in the limit $v \ll c$, because the basic equation of particle dynamics (the 2nd Newton law)¹

$$m_k \ddot{\mathbf{r}}_k = -\nabla_k \sum_{k'} U(\mathbf{r}_k - \mathbf{r}_{k'}), \quad (9.1)$$

where U , the potential energy of inter-particle interactions, is invariant with respect to the so-called *Galilean transform* (or “transformation”).² Choosing the coordinate axes of both frames so that axes x and x' are parallel to vector \mathbf{v} (Fig. 1), the transform³ may be presented as

Galilean
transform

$$x = x' + vt', \quad y = y', \quad z = z', \quad t = t', \quad (9.2a)$$

¹ Let me hope that the reader does not need a reminder that in order for Eq. (1) to be valid, the reference frames 0 and 0' have to be inertial – see, e.g., CM Sec. 1.3.

² It had been first formulated by G. Galilei as early as in 1638 – four years before I. Newton was *born*!

³ Note a very unfortunate term “boost”, used sometimes for the transform between the reference frames. (It is especially unnatural in the special relativity, not describing any accelerations.) In these notes, this term is avoided.

and plugging Eq. (2a) into Eq. (1), we get an absolutely similarly looking equation of motion in the “moving” reference frame $0'$. Since the reciprocal transform,

$$x' = x - vt, \quad y = y', \quad z' = z, \quad t' = t, \quad (9.2b)$$

is similar to the direct one, with the replacement of $(+v)$ with $(-v)$, we may say that the Galilean invariance means that there is no any “master” (*absolute*) spatial reference frame in classical mechanics, although the spatial and temporal intervals between different events are absolute (reference-frame invariant).

However, it is straightforward to use Eq. (2) to check that the form of the wave equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \right) f = 0, \quad (9.3)$$

describing in particular the electromagnetic wave propagation in free space,⁴ is *not* Galilean-invariant.⁵ For the “usual” (say, elastic) waves, which obey a similar equation albeit with a different speed,⁶ this lack of Galilean invariance is natural and is compatible with the invariance of Eq. (1), from which the wave equation originates. This is because the elastic waves are essentially the oscillations of interacting particles of a certain medium (e.g., an elastic solid), which makes the reference frame connected to this medium, special. So, if the electromagnetic waves were oscillations of a certain special medium (that was first called the “luminiferous aether”⁷ and later just *ether*), similar arguments might be applicable to reconcile Eqs. (2) and (3).

The detection of such a medium was the goal of the Michelson-Morley measurements (carried out between 1881 and 1887 with better and better precision), that are sometimes called “the most famous failed experiment in physics”. Figure 2 shows a crude scheme of their experiments.

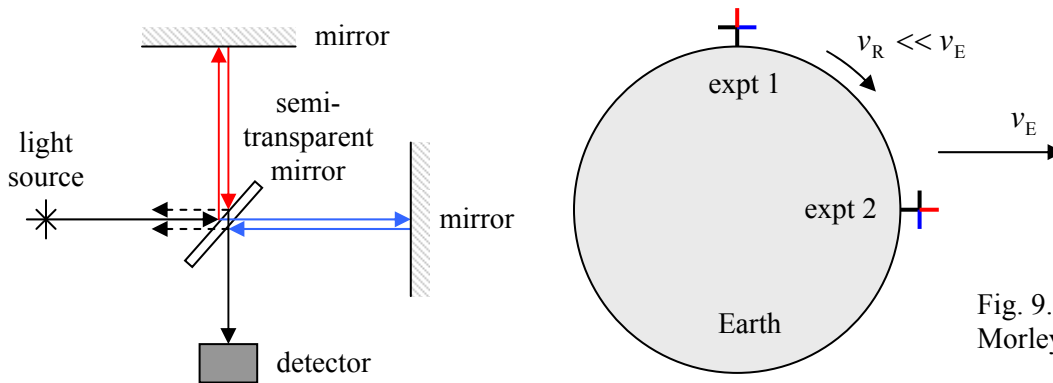


Fig. 9.2. The Michelson-Morley experiment.

⁴ Discussions in this chapter and most of the next chapter will be restricted to the free-space (and hence dispersion-free) case; some media effects on radiation by relativistic particles will be discussed in Sec.10.4.

⁵ It is interesting that the Schrödinger equation, whose fundamental solution for a free particle is a similar monochromatic wave (albeit with a different dispersion law), is Galilean-invariant, with a certain addition to the wavefunction's phase – see, e.g., QM Chapter 1.

⁶ See, e.g., CM Secs. 5.5 and 7.7.

⁷ In the ancient Greek mythology, aether is the clear upper air breathed by the gods residing on mount Olympus.

A nearly-monochromatic wave is split in two parts (nominally, of equal intensity), using a semi-transparent mirror tilted by 45° to the incident wave direction. These two partial waves are reflected back by two genuine mirrors, and arrive at the same semi-transparent mirror again. Here a half of each wave is returned to the light source area (where they vanish without affecting the source), but another half passes toward the detector, forming, with its counterpart, an interference pattern similar to that in the Young experiment. Thus each of the interfering waves has traveled twice (back and forth) each of two mutually perpendicular “arms” of the interferometer. Assuming that the ether, in which light propagates with speed c , moves with speed $v < c$ along one of the arms, of length l_t , it is straightforward (and hence left for reader’s exercise :-)) to get the following expression for the difference between light roundtrip times:

$$\Delta t = \frac{2}{c} \left(\frac{l_t}{(1 - v^2/c^2)^{1/2}} - \frac{l_t}{1 - v^2/c^2} \right) \approx \frac{l}{c} \left(\frac{v}{c} \right)^2, \quad (9.4)$$

where l_t is the length of the second, “transverse” arm of the interferometer (perpendicular to \mathbf{v}), and the last, approximate expression is valid at $l_t \approx l_l$ and $v \ll c$.

Since Earth moves around the Sun with speed $v_E \approx 30 \text{ km/s} \approx 10^{-4} c$, the arm positions relative to this motion alternate, due to Earth rotation about its axis, each 6 hours – see the right panel of Fig. 2. Hence if we assume that the ether rests in Sun’s reference frame, Δt (and the corresponding shift of interference fringes), has to alternate with this half-period as well. The same alternation may be achieved, at a smaller time scale, by a deliberate rotation of the instrument by $\pi/2$. In the most precise version of the Michelson-Morley experiment (1887), this shift was expected to be close to 0.4 of the fringe pattern period. The result was negative, with the error bar about 0.01 of the fringe period.⁸

The most prominent immediate explanation of this zero result⁹ was suggested in 1889 by G. FitzGerald and (independently and more qualitatively) by H. Lorentz in 1892: as evident from Eq. (4), if the longitudinal arm of the interferometer itself experiences the so-called *length contraction*,

$$l_l(v) = l_l(0) \left(1 - \frac{v^2}{c^2} \right)^{1/2}, \quad (9.5)$$

while the transverse arm’s length is not affected by the motion relative to the ether, this kills Δt . This, extremely radical, idea received a strong support from the proof, in 1887-1905, that the Maxwell equations, and hence the wave equation (3), are form-invariant under the so-called *Lorentz transform*.¹⁰ For the choice of coordinates shown in Fig. 1, the transform reads

⁸ Through the 20th century, the Michelson-Morley-type experiments were repeated using more and more refined experimental techniques, always with the zero result for the apparent ether motion speed. For example, recent experiments, using cryogenically cooled optical resonators, have reduced the upper limit for such speed to just $3 \times 10^{-15} c$ – see H. Müller *et al.*, *Phys. Rev. Lett.* **91**, 020401 (2003).

⁹ The zero result of a slightly later experiment, namely precise measurements of the torque which should be exerted by the moving ether on a charged capacitor, carried out in 1903 by F. Trouton and H. Noble (following G. FitzGerald’s suggestion), seconded the Michelson and Morley’s conclusions.

¹⁰ The theoretical work toward this goal (which I do not have time to review in detail) included important contributions by W. Voigt (in 1887), H. Lorentz (1892 - 1904), J. Larmor (1897 and 1900), and H. Poincaré (1900 and 1905).

$$x = \frac{x' + vt'}{(1 - v^2/c^2)^{1/2}}, \quad y = y', \quad z = z', \quad t = \frac{t' + (v/c^2)x'}{(1 - v^2/c^2)^{1/2}}. \quad (9.6a) \quad \text{Lorentz transform}$$

It is elementary to solve these equations for the primed coordinates to get the reciprocal transform

$$x' = \frac{x - vt}{(1 - v^2/c^2)^{1/2}}, \quad y' = y, \quad z' = z, \quad t' = \frac{t - (v/c^2)x}{(1 - v^2/c^2)^{1/2}}. \quad (9.6b)$$

(I will soon present Eqs. (6) in a more elegant form.)

The Lorentz transform relations (6) are evidently reduced to the Galilean transform formulas (2) at $v^2 \ll c^2$. As will be proved in the next section, Eqs. (6) also yield the Lorentz length contraction (5). However, all attempts to give a reasonable interpretation of these equations while keeping the notion of the ether have failed, in particular because of the restrictions imposed by results of earlier experiments carried out in 1851 and 1853 by H. Fizeau - that were repeated with higher accuracy by the same Michelson and Morley in 1886. These experiments have shown that if one sticks to the ether concept, this hypothetical medium should be partially “dragged” by any moving dielectric media with a speed proportional to $(\epsilon_r - 1)$. Careful reasoning shows that such local drag is irreconcilable with the assumed continuity of the ether.

In his famous 1905 paper, Albert Einstein has suggested a bold resolution of this contradiction, essentially removing the concept of the ether altogether. Moreover, he argued that the Lorentz transform is the general property of time and space, rather than of the electromagnetic field alone. He has started with two postulates, the first one essentially repeating the principle of relativity, formulated earlier (1904) by H. Poincaré in the following form:

“...the laws of physical phenomena should be the same, whether for an observer fixed, or for an observer carried along in a uniform movement of translation; so that we have not and could not have any means of discerning whether or not we are carried along in such a motion.”¹¹

The second Einstein’s postulate was that the speed of light c , in free space, should be constant in all reference frames. (This is essentially a denial of ether’s existence.)

Then, Einstein showed how naturally do the Lorentz transform relations (6) follow from his postulates, with a few (very natural) additional assumptions. Let a point source emit a short flash of light, at the moment $t = t' = 0$ when origins of the reference frames shown in Fig. 1 coincide. Then, according to the second of Einstein’s postulates, in each of the frames the spherical wave propagates with the same speed c , i.e. coordinates of points of its front, measured in the two frames, have to obey equations

$$\begin{aligned} (ct)^2 - (x^2 + y^2 + z^2) &= 0, \\ (ct')^2 - (x'^2 + y'^2 + z'^2) &= 0. \end{aligned} \quad (9.7)$$

What may be the general relation between the combinations in the left-hand side of these equations - not for this selected pair of events, the light flash and its detection, but in general? A very natural (essentially, the only justifiable) choice is

¹¹ Note that though the relativity principle excludes the notion of the special (“absolute”) spatial reference frame, its verbal formulation still leaves the possibility of the Galilean “absolute time” open. The quantitative relativity theory kills this option – see Eqs. (6) and their discussion below.

$$[(ct)^2 - (x^2 + y^2 + z^2)] = f(v^2)[(ct')^2 - (x'^2 + y'^2 + z'^2)]. \quad (9.8)$$

Now, according to the first postulate, the same relation should be valid if we swap the reference frames ($x \leftrightarrow x'$, etc.)¹² and replace v with $(-v)$. This is only possible if $f^2 = 1$, so that excluding option $f = -1$ (which is incompatible with the Galilean transform in the limit $v/c \rightarrow 0$), we get

$$(ct)^2 - (x^2 + y^2 + z^2) = (ct')^2 - (x'^2 + y'^2 + z'^2). \quad (9.9)$$

For the line $y = y' = 0, z = z' = 0$, Eq. (9) is reduced to

$$(ct)^2 - x^2 = (ct')^2 - x'^2. \quad (9.10)$$

It is very illuminating to interpret this relation as the one resulting from a mutual rotation of the reference frames (that now have to include clocks to measure time) on the plane of coordinate x and the so-called *Euclidian time* $\tau \equiv ict$ – see Fig. 3.

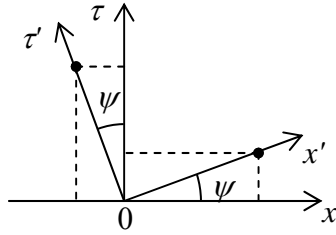


Fig. 9.3. The Lorentz transform as a mutual rotation of reference frames on the $[x, \tau]$ plane.

Indeed, rewriting Eq. (10) as

$$\tau^2 + x^2 = \tau'^2 + x'^2, \quad (9.11)$$

we may consider it as the invariance of the squared radius at the rotation that is shown in Fig. 3 and described by the evident geometric relations

$$\begin{aligned} x &= x' \cos \psi - \tau' \sin \psi, \\ \tau &= x' \sin \psi + \tau' \cos \psi, \end{aligned} \quad (9.12a)$$

with the reciprocal relations

$$\begin{aligned} x' &= x \cos \psi + \tau \sin \psi, \\ \tau' &= -x \sin \psi + \tau \cos \psi. \end{aligned} \quad (9.12b)$$

So far, angle ψ has been arbitrary. In the spirit of Eq. (8), a natural choice is $\psi = \psi(v)$, with the requirement $\psi(0) = 0$. In order to find this function, let us write the definition of velocity v of frame $0'$, as measured in reference frame 0 : for $x' = 0, x = vt$. In variables x and τ , this means

$$\left. \frac{x}{\tau} \right|_{x'=0} \equiv \left. \frac{x}{ict} \right|_{x'=0} = \frac{v}{ic}. \quad (9.13)$$

On the other hand, for the same point $x' = 0$, Eqs. (12a) yield

¹² Strictly speaking, at this swap we should also replace v with $(-v)$, but this change does not affect Eq. (8).

$$\frac{x}{\tau} \Big|_{x'=0} = -\tan \psi. \quad (9.14)$$

These two expressions are compatible only if

$$\tan \psi = \frac{iv}{c}, \quad (9.15)$$

so that

$$\sin \psi \equiv \frac{\tan \psi}{(1 + \tan^2 \psi)^{1/2}} = \frac{iv/c}{(1 - v^2/c^2)^{1/2}} \equiv i\beta\gamma, \quad \cos \psi \equiv \frac{1}{(1 + \tan^2 \psi)^{1/2}} = \frac{1}{(1 - v^2/c^2)^{1/2}} \equiv \gamma, \quad (9.16)$$

where β and γ are two very convenient and commonly used dimensionless parameters defined as

$$\beta \equiv \frac{\mathbf{v}}{c}, \quad \gamma \equiv \frac{1}{(1 - v^2/c^2)^{1/2}} = \frac{1}{(1 - \beta^2)^{1/2}}. \quad (9.17) \quad \text{Parameters } \beta \text{ and } \gamma$$

(Vector β is called the *normalized velocity*, while scalar γ , the *Lorentz factor*.)¹³

Using the relations for ψ , Eqs. (12) become

$$x = \gamma(x' - i\beta\tau'), \quad \tau = \gamma(i\beta x' + \tau'), \quad (9.18a)$$

$$x' = \gamma(x + i\beta\tau), \quad \tau' = \gamma(-i\beta x + \tau). \quad (9.18b)$$

Now returning to the real variables $[x, ct]$, we get the Lorentz transform relations (6) in a more compact form:¹⁴

$$x = \gamma(x' + \beta ct'), \quad y = y', \quad z = z', \quad ct = \gamma(ct' + \beta x'), \quad (9.19a)$$

$$x' = \gamma(x - \beta ct), \quad y' = y, \quad z' = z, \quad ct' = \gamma(ct - \beta x). \quad (9.19b)$$

An immediate corollary of Eqs. (6) is that for γ to stay real, we need $v^2 \leq c^2$, i.e. that the speed of any physical body (to which we could connect a reference frame) cannot exceed the speed of light, as measured in *any* physically meaningful reference frame.¹⁵

9.2. Relativistic kinematic effects

In order to discuss other corollaries of Eqs. (19), we need to spend a few minutes to discuss what do these relations actually mean. Evidently, they are trying to tell us that the spatial and temporal intervals are not absolute (as they are in the Newtonian space), but do depend on the reference frame they are measured in. So, we have to understand very clearly what exactly may be measured - and thus may be discussed in a physics theory. Recognizing this necessity, A. Einstein has introduced the notion of numerous imaginary *observers* that may be distributed all over each reference frame. Each observer has a clock and may use it to measure the instants of *local* events. He also conjectured that:

¹³ One more function of β , the *rapidity* defined as $\beta \equiv \tanh \phi$ (so that $\psi = i\phi$), is also useful for some calculations.

¹⁴ Still, in some cases below, it will be more convenient to use Eqs. (6) rather than Eqs. (19).

¹⁵ All attempts to rationally conjecture particles moving with $v > c$, called *tachyons*, have failed (so far, at least :-). Possibly the strongest objection against their existence is the notice that tachyons could be used to communicate back in time, thus violating the causality principle – see, e.g., G. Benford *et al.*, *Phys. Rev. D* **2**, 263 (1970).

(i) all observers within the same reference frame may agree on a common length measure (“a scale”), i.e. on their relative positions in that frame, and synchronize their clocks,¹⁶ and

(ii) observers belonging to different reference frames may agree on the nomenclature of *world events* (e.g., short flashes of light) to which their respective measurements belong.

Actually, these additional postulates have been already implied in our “derivation” of the Lorentz transform in Sec. 1. For example, by $\{x, y, z, \text{ and } t\}$ we mean the results of space and time measurements of a certain world event, about that all observers belonging to frame 0 agree. Similarly, all observers of frame 0' have to agree about results $\{x', y', z', t'\}$. Finally, when the origin of frame 0' passes by some sequential points x_k of frame 0, observers in that frame may measure its passage times t_k without a fundamental error, and know that all these times belong to $x' = 0$.

Now we can analyze the major corollaries of the Lorentz transform, which are rather striking from the point of view of our everyday (rather nonrelativistic :-) experience.

(i) Length contraction. Let us consider a rigid rod, stretched along axis x , with length $l = x_2 - x_1$, where $x_{1,2}$ are the coordinates of rod's ends, as measured in its rest frame 0, at any instant t (Fig. 4). What would be the rod's length l' measured by the Einstein observers in the moving frame 0'?

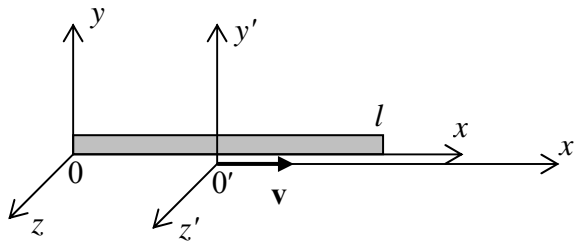


Fig. 9.4. Relativistic length contraction.

At a time instant t' agreed upon in advance, the observers who find themselves exactly at the rod's ends, may register that fact, and then subtract their coordinates $x'_{1,2}$ to calculate the apparent rod length $l' = x'_2 - x'_1$ in the moving frame. According to Eq. (19a), l may be expressed via l' as

$$l \equiv x_2 - x_1 = \gamma(x'_2 + \beta ct') - \gamma(x'_1 + \beta ct') = \gamma(x'_2 - x'_1) = \gamma l' > l'. \quad (9.20a)$$

Hence, the rod's length, as measured in the *moving* reference frame is

Length
contraction

$$l' = \frac{l}{\gamma} = l \left(1 - \frac{v^2}{c^2} \right)^{1/2} \leq l, \quad (9.20b)$$

in accordance with the FitzGerald-Lorentz hypothesis (5). This is the *relativistic length contraction* effect: an object is always the longest (has the so-called *proper length* l) if measured in its *rest frame*. Note that according to Eq. (19), the length contraction takes place only in the direction of the relative motion of two reference frames. As has been noted in Sec. 1, this result immediately explains the zero

¹⁶ A posteriori, the Lorentz transform may be used to show that consensus-creating procedures (such as clock synchronization) are indeed possible. The basic idea of the proof is that at $v \ll c$ the relativistic corrections to space and time intervals are of the order of $(v/c)^2$, they have negligible effects on clocks being brought together into the same point for synchronization very slowly, with velocity $v \ll c$. The reader interested in detailed discussion of this and other fine points of special relativity may be referred to, e.g., either H. Arzeliès, *Relativistic Kinematics*, Pergamon, 1966, or W. Rindler, *Introduction to Special Relativity*, 2nd ed., Oxford U. Press, 1991.

result of the Michelson-Morley-type experiments, so that they give a convincing evidence (if not an irrefutable proof) of Eq. (20).

(ii) Time dilation. Now let us use Eqs. (19a) to find the time interval Δt , as measured in frame 0, between two world events – say, two ticks of a clock moving with frame 0' (Fig. 5), i.e. having constant values of x' , y' , and z' .

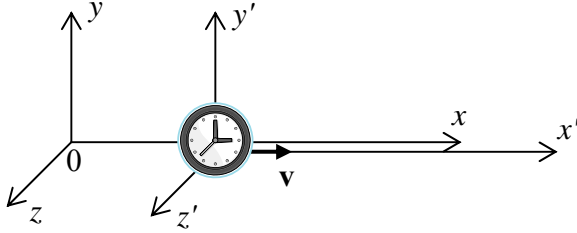


Fig. 9.5. Relativistic time dilation.

Let the time interval between these two events, measured in clock's rest frame 0', be $\Delta t' \equiv t_2' - t_1'$. At these two moments, the clock would fly by certain two Einstein's observers at rest in frame 0, so that they can record the corresponding moments $t_{1,2}$ shown by their clocks, and then calculate Δt as their difference. According to the second of Eqs. (19a),

$$\Delta t \equiv t_2 - t_1 = \frac{\gamma}{c} [(ct_2' + \beta x') - (ct_1' + \beta x')] = \gamma \Delta t', \quad (9.21a)$$

so that, finally,

$$\Delta t = \gamma \Delta t' \equiv \frac{\Delta t'}{(1 - v^2/c^2)^{1/2}} \geq \Delta t'. \quad (9.21b) \quad \text{Length contraction}$$

This is the famous *relativistic time dilation* (or “dilatation”) effect: a time interval is *longer* if measured in a frame (in our case, frame 0) *moving relatively to the clock*, while that in the rest frame is the shortest - the so-called *proper time interval*.

This counter-intuitive effect is the everyday reality at experiments with high-energy elementary particles. For example, in a typical (by no means record-breaking) experiment carried out in Fermilab, a beam of charged 200 GeV pions with $\gamma \approx 1,400$ passed distance $l = 300$ m distance with the measured loss of only 3% of the initial beam intensity due to the pion decay (mostly, into muon-neutrino pairs) with proper lifetime $t_0 \approx 2.56 \times 10^{-8}$ s. Without the time dilation, only an $\exp\{-l/ct_0\} \sim 10^{-17}$ part of the initial pions would survive, while the relativity-corrected number $\exp\{-l/ct\} = \exp\{-l/c\gamma t_0\} \approx 0.97$ was in a full accordance with experimental measurements. As another example, the global positioning system (GPS) is designed with the account of the time dilation due to the velocity of its satellites (and also some gravity-induced, i.e. general-relativity corrections that I do not have time to discuss) and would give large errors without such corrections. So, there is no doubt that time dilation (21) is a reality, though the precision of its experimental tests I am aware of has been limited by a few percent, because of almost unavoidable involvement of gravity effects.¹⁷

Before the first reliable observation of the time dilation (by B. Rossi and D. Hall in 1940), there had been serious doubts in the reality of this effect, the most famous being the *twin paradox* first posed

¹⁷ See, e.g., J. Hafele and R. Keating, *Science* **177**, 166 (1972).

(together with an immediate suggestion of its resolution) by P. Langevin in 1911. Let us send one of two twins on a long star journey with a speed v approaching c . Upon his return to Earth, who of the twins would be older? The naïve approach is to say that due to the relativity principle, not one can be (and hence there is no time dilation), because each twin could claim that his counterpart, rather than himself, was moving, with the same speed v , just in the opposite direction. The resolution of the paradox in the general theory of relativity (which can handle gravity and acceleration effects) is that one of the twins had to be accelerated to be brought back, and hence the reference frames have to be dissimilar: only one of them may stay inertial all the time. Because of that, the twin who had been accelerated (“actually traveling”) would be younger than his sibling when they meet.

(iii) Velocity transformation. Now let us calculate velocity \mathbf{u} of a particle, as observed in reference frame 0, provided that its velocity, as measured in frame 0', is \mathbf{u}' (Fig. 6).

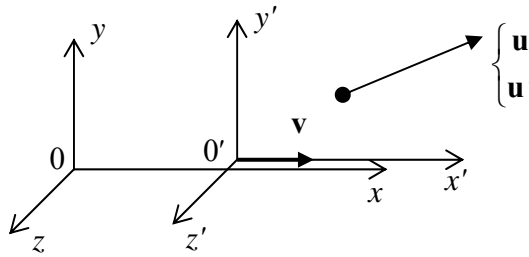


Fig. 9.6. Relativistic velocity addition.

Keeping the usual definition of velocity, but with due attention to the relativity of not only spatial but also temporal intervals, we may write

$$\mathbf{u} \equiv \frac{d\mathbf{r}}{dt}, \quad \mathbf{u}' \equiv \frac{d\mathbf{r}'}{dt'}. \quad (9.21)$$

Plugging in the differentials of the Lorentz transform relations (6a), we get

$$u_x = \frac{dx}{dt} = \frac{dx' + v dt'}{dt' + v dx' / c^2} = \frac{u'_x + v}{1 + u'_x v / c^2}, \quad u_y = \frac{dy}{dt} = \frac{1}{\gamma} \frac{dy'}{dt' + v dx' / c^2} = \frac{1}{\gamma} \frac{u'_y}{1 + u'_x v / c^2}, \quad (9.22)$$

and the similar formula for u_z . In the classical limit $v/c \rightarrow 0$, these relations are reduced to

$$u_x = u'_x + v, \quad u_y = u'_y, \quad u_z = u'_z, \quad (9.23)$$

and may be merged into the familiar Galilean vector form

$$\mathbf{u} = \mathbf{u}' + \mathbf{v}, \quad \text{for } v \ll c. \quad (9.24)$$

In order to see how strange the full relativistic rules (22) are, let us first consider a purely longitudinal motion, $u_y = u_z = 0$; then¹⁸

$$u = \frac{u' + v}{1 + u'v / c^2}, \quad (9.25)$$

Longitudinal
velocity
addition

¹⁸ With an account of the well-known trigonometric identity $\tan(a + b) = (\tan a + \tan b) / (1 - \tan a \tan b)$ and Eq. (15), Eq. (25) shows that that rapidities ψ add up exactly as longitudinal velocities at nonrelativistic motion, making that notion very convenient for the analysis transfer between several frames.

where $u \equiv u_x$ and $u' \equiv u'_x$. Figure 7 shows u as the function of u' , given by this formula, for several values of the reference frames' relative velocity v .

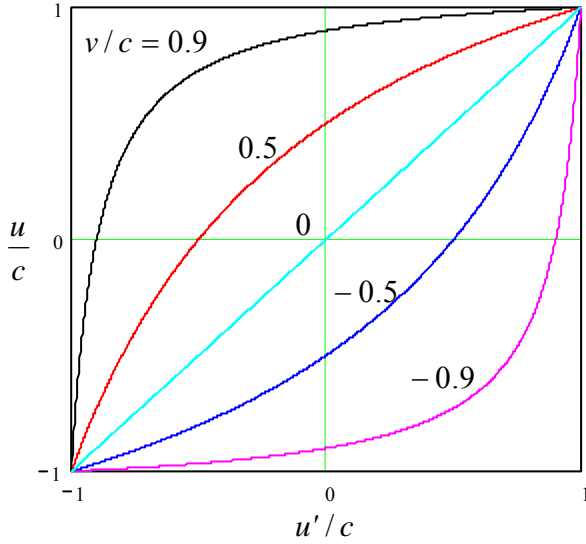


Fig. 9.7. Longitudinal velocity addition.

The first sanity check is that if $v = 0$, i.e. the reference frames are at rest relative to each other, then $u = u'$, as it should be – see the diagonal straight line. Next, if magnitudes of u' and v are both below c , so is the magnitude of u . (Also good, because otherwise ordinary particles in one frame would be tachyons in the other one, and the theory would be in a big trouble.) Now strange things start: even as u' and v are both approaching c , then u is also close to c , but does not exceed it. As an example, if we fired ahead a bullet with speed $0.9c$ from a spaceship moving from the Earth also at $0.9c$, Eq. (25) predicts the speed of the bullet relative to Earth to be just $[(0.9 + 0.9)/(1 + 0.9 \times 0.9)]c \approx 0.994c < c$, rather than $(0.9 + 0.9)c = 1.8c > c$ as in the Galilean kinematics. We certainly should accept this strangeness of relativity, because it is necessary to fulfill the 2nd Einstein's postulate: the independence of the speed of light in any reference frame. Indeed, for $u' = \pm c$, Eq. (25) yields $u = \pm c$, regardless of v .

In the opposite case of transverse motion, when a particle moves across the relative motion of the frames (for example, at our choice of coordinates, $u'_x = u'_z = 0$), Eqs. (22) yield a less spectacular result

$$u_y = \frac{u'_y}{\gamma} \leq u'_y. \quad (9.26)$$

This effect comes purely from the time dilation, because the transverse coordinates are Lorentz-invariant.

In the case when both u'_x and u'_y are substantial (but u'_z is still zero), we may divide expressions (22) by each other to relate angles θ of particle propagation, as observed in the two reference frames:

$$\tan \theta \equiv \frac{u_y}{u_x} = \frac{u'_y}{\gamma(u'_x + v)} = \frac{\sin \theta'}{\gamma(\cos \theta' + v/u')}. \quad (9.27)$$

Stellar
aberration
effect

This expression describes, in particular, the so-called *stellar aberration* effect, the dependence of the observed direction θ toward a star on the speed v of the telescope motion relative to the star – see Fig. 8. (The effect is readily observable experimentally as the *annual aberration* due to the periodic change

of speed v by $2v_E \approx 60$ km/s because of Earth's rotation about the Sun. Since the aberration's main part is of the first order in $v_E/c \sim 10^{-4}$, the effect is very significant and has been known since the early 1700s.)

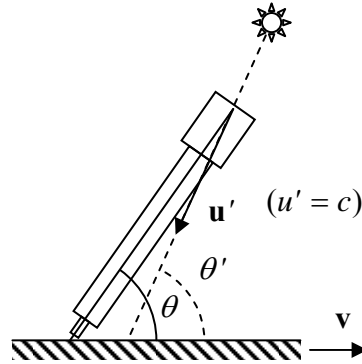


Fig. 9.8. Stellar aberration.

For the analysis of this effect, it is sufficient to take, in Eq. (27), $u' = c$, i.e. $v/u' = \beta$, and interpret θ' as the “proper” direction to the star, that would be measured at $v = 0$.¹⁹ At $\beta \ll 1$, both Eq. (27) and the Galilean result (which the reader is invited to derive directly from Fig. 8),

$$\tan \theta = \frac{\sin \theta'}{\cos \theta' + \beta}, \quad (9.28)$$

may be well approximated by the first-order term

$$\Delta \theta \equiv \theta' - \theta \approx \beta \sin \theta. \quad (9.29)$$

Unfortunately, it is not easy to use the difference between Eqs. (27) and (28), of the second order in β , for the special relativity confirmation, because other components of Earth's motion, such as its rotation, nutation and torque-induced precession,²⁰ give masking first-order contributions to the aberration.

Finally, at a completely arbitrary direction of vector \mathbf{u}' , Eqs. (22) may be readily used to calculate the velocity magnitude. The most popular form of the resulting expression is for the square of relative velocity (or rather relative reduced velocity β) of two particles,

$$\beta^2 = \frac{(\beta_1 - \beta_2)^2 - |\beta_1 \cdot \beta_2|}{(1 - \beta_1 \cdot \beta_2)^2} \leq 1. \quad (9.30)$$

where $\beta_{1,2} \equiv \mathbf{v}_{1,2}/c$ are their normalized velocities as measured in the same reference frame.

(iv) The Doppler effect. Now let us consider a plane, monochromatic wave moving along axis x :

$$f = \text{Re}[f_\omega \exp\{i(kx - \omega t)\}] = |f_\omega| \cos(kx - \omega t + \arg f_\omega). \quad (9.31)$$

¹⁹ Strictly speaking, in order to reconcile the geometries shown in Fig. 1 (for which all our formulas, including Eq. (27), are valid) and Fig. 8 (giving the traditional scheme of the aberration), it is necessary to invert signs of \mathbf{u} (and hence $\sin \theta'$ and $\cos \theta'$) and \mathbf{v} , but as evident from Eq. (27), all the minus signs cancel, and the formula is valid as is.

²⁰ See, e.g., CM Secs. 6.4-6.5.

Its total phase, $\Psi \equiv kx - \omega t + \arg f_\omega$ (in contrast to amplitude $|f_\omega|$!) cannot depend on the observer's reference frame, because all fields of a traveling wave vanish simultaneously at $\Psi = 2\pi n$, (for all integer n) and such “world events” should take place in all reference frames. The only way to keep $\Psi = \Psi'$ at all times is to have²¹

$$kx - \omega t = k'x' - \omega't' . \quad (9.32)$$

First, let us consider the Doppler effect describing usual nonrelativistic waves, e.g., oscillations of particles of a certain medium. Using the Galilean transform (2), we may rewrite Eq. (32) as

$$k(x' + vt) - \omega t = k'x' - \omega't' . \quad (9.33)$$

Since this transform leaves all space intervals (including wavelength $\lambda = 2\pi/k$) intact, we can take $k = k'$, so that Eq. (33) yields

$$\omega' = \omega - kv . \quad (9.34)$$

For a dispersion-free medium, the wave number k is the ratio of its frequency ω , as measured in the reference frame bound to the medium, and the wave velocity v_w . In particular, if the wave source rests in the medium, we can bind frame 0 to the medium as well, and frame 0' to wave's receiver (so that $v = v_r$), so that

$$k = \frac{\omega}{v_w} , \quad (9.35)$$

and for the frequency perceived by the receiver, Eq. (34) yields

$$\omega' = \omega \frac{v_w - v_r}{v_w} . \quad (9.36)$$

On the other hand, if the receiver and the medium are at rest in reference frame 0', while the wave source is bound to frame 0 (so that $v = -v_s$), Eq. (35) should be replaced with

$$k = k' = \frac{\omega'}{v_w} , \quad (9.37)$$

and Eq. (34) yields a different result:

$$\omega' = \omega \frac{v_w}{v_w - v_s} , \quad (9.38)$$

Finally, if both the source and detector are moving, it is straightforward to combine these two results to get the general relation

$$\omega' = \omega \frac{v_w - v_r}{v_w - v_s} . \quad (9.39)$$

At low speeds of both the source and receiver, this result simplifies,

²¹ Strictly speaking, Eq. (32) is valid to an additive constant, but for notation simplicity, it may be always made equal to zero by selecting (at it has already been done in all relations of Sec. 1) the reference frame origins and/or clock turn-on times so that at $t = 0$ and $x = 0$, $t' = 0$ and $x' = 0$ as well.

$$\omega' \approx \omega(1 - \beta), \quad \beta \equiv \frac{v_r - v_s}{v_w}, \quad (9.40)$$

but at speeds comparable to v_w we have to use the more general Eq. (39). Thus, the usual Doppler effect is affected not only by the relative speed ($v_r - v_s$) of wave's source and detector, but also of their speeds relative to the medium in which waves propagate.

Somewhat counter-intuitively, for the electromagnetic waves the calculations are simpler, because for them the propagation medium (ether) does not exist, wave velocity equals $\pm c$ in any reference frame, and there are no two separate cases: we can always take $k = \pm \omega/c$, $k' = \pm \omega'/c$. Plugging these relations, together with the Lorentz transform (19a), into the phase-invariance equation (32), we get

$$\pm \frac{\omega}{c} \gamma(x' + \beta ct') - \omega \gamma \frac{ct' + \beta x'}{c} = \pm \frac{\omega'}{c} x' - \omega' t'. \quad (9.41)$$

This relation has to hold for any x' and t' , so we may require the net coefficients before these variables to vanish. These two requirements yield the same condition:

$$\omega' = \omega \gamma(1 \mp \beta). \quad (9.42)$$

This result is already quite simple, but may be transformed further to be even more illuminating:

$$\omega' = \omega \frac{1 \mp \beta}{(1 - \beta^2)^{1/2}} = \omega \left[\frac{(1 \mp \beta)(1 \mp \beta)}{(1 + \beta)(1 - \beta)} \right]^{1/2}. \quad (9.43)$$

At any sign before β , one pair of parentheses cancel, so that

Longitudinal
Doppler
effect

$$\omega' = \omega \left(\frac{1 \mp \beta}{1 \pm \beta} \right)^{1/2}. \quad (9.44)$$

(It may look like the reciprocal expression of ω via ω' is different, violating the relativity principle. However, in this case we have to change the sign of β , because the relative velocity of the system is opposite, so we come down to Eq. (44) again.)

Thus the Doppler effect for electromagnetic waves depends only on the relative velocity $v = \beta c$ between the wave source and detector – as it should be, given the absence of the ether. At velocities much below c , Eq. (43) may be crudely approximated as

$$\omega' \approx \omega \frac{1 \mp \beta/2}{1 \pm \beta/2} \approx \omega(1 \mp \beta), \quad (9.45)$$

i.e. in the first approximation in $\beta \equiv v/c$ it coincides with the corresponding limit (38) of the usual Doppler effect. However, even at $v \ll c$ there is still a difference of the order of $(v/c)^2$ between the Galilean and Lorentzian relations.

If the wave vector \mathbf{k} is tilted by angle θ to vector \mathbf{v} (as measured in frame 0), then we have to repeat the calculations, with k replaced by k_x , and components k_y and k_z left intact at the Lorentz transform. As a result, Eq. (42) is generalized as

$$\omega' = \omega \gamma(1 - \beta \cos \theta). \quad (9.46)$$

For the cases $\cos\theta = \pm 1$, Eq. (44) reduces to our previous result. However, at $\theta = \pi/2$ (i.e. $\cos\theta = 0$), the relation is rather different:

$$\omega' = \gamma\omega = \frac{\omega}{(1 - \beta^2)^{1/2}}. \quad (9.47) \quad \text{Transverse Doppler effect}$$

This is the *transverse Doppler effect* - which is completely absent in the nonrelativistic physics. Its first experimental evidence was obtained using electron beams (as suggested in 1906 by J. Stark), by H. Ives and G. Stilwell in 1938 and 1941. Later, similar experiments were repeated several times, but the first unambiguous measurement were performed only in 1979 by D. Hasselkamp *et al.* who confirmed Eq. (47) with a relative accuracy about 10%. This precision may not look too spectacular, but besides the special tests discussed above, the Lorentz transform formulas have been also confirmed, less directly, by a huge body of other experimental data, especially in high energy physics, being in agreement with calculations incorporating the transform as their part. This is why, with every respect to the challenging authority spirit, I should warn the reader: you decide to challenge the relativity theory (that is called “theory” by tradition only), you would also need to explain all these data.²² Best luck with that!

9.3. 4-vectors, momentum, mass, and energy

Before proceeding to relativistic dynamics, let us discuss a mathematical language that makes all the calculations more compact - and more beautiful. We have already seen that spatial coordinates $\{x, y, z\}$ and product ct are Lorentz-transformed similarly – see Eqs. (19). So it is natural to consider them as components of a 4-component vector (or, for short, *4-vector*),

$$\{x_0, x_1, x_2, x_3\} = \{ct, \mathbf{r}\}, \quad (9.48)$$

with components

$$x_0 = ct, \quad x_1 = x, \quad x_2 = y, \quad x_3 = z. \quad (9.49) \quad \text{Space-time 4-vector}$$

According to Eqs. (19), its components are Lorentz-transformed as

$$x_j = \sum_{j'=0}^3 L_{jj'} x'_{j'}, \quad (9.50) \quad \text{4-form of Lorentz Transform}$$

where $L_{jj'}$ are the elements of the 4×4 Lorentz transform matrix

$$\begin{pmatrix} \gamma & \beta\gamma & 0 & 0 \\ \beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (9.51) \quad \text{Lorentz transform matrix}$$

Since 4-vectors are a new notion for our course, and are used for much more goals than the just the space-time transform, we need to discuss the mathematical rules they obey. Indeed, as was

²² The same fact, ignored by crackpots, is also valid for other favorite points of their attacks, including the Universe expansion and quantum mechanics in physics, and the evolution theory in biology.

mentioned in Sec. 8.9, the usual (3D) vector is not just any ordered set (*string*) of three scalars $\{A_x, A_y, A_z\}$; if we want it to represent a reference-frame-independent physical variable, vector components have to obey certain rules at transfer from one reference frame to another. In particular, vector's 3D *norm* (magnitude squared),

$$A^2 = A_x^2 + A_y^2 + A_z^2, \quad (9.52)$$

should be an invariant at the Galilean transform (2). However, a naïve extension of this formula to 4-vectors would not work, because, according to the calculations of Sec. 1, the Lorentz transform keeps intact combinations of the type (7), with one sign negative, rather than the sum of all components squared. Hence for the 4-vector all the rules of the game have to be reviewed and adjusted – or rather redefined from the very beginning.

Arbitrary 4-vector is a string of 4 scalars,

General
4-vector

$$\{A_0, A_1, A_2, A_3\}, \quad (9.53)$$

defined in 4D *Minkowski space*,²³ whose components A_j , as measured in systems 0 and 0', shown in Fig. 1, obey the Lorentz transform similar to Eq. (50):

General
4-vector's
Lorentz
transform

$$A_j = \sum_{j'=0}^3 L_{jj'} A'_{j'}. \quad (9.54)$$

As we have already seen on the example of the space-time 4-vector (48), this means in particular that

Lorentz
invariance

$$A_0^2 - \sum_{j=1}^3 A_j^2 = (A'_0)^2 - \sum_{j=1}^3 (A'_j)^2. \quad (9.55)$$

This is the so-called *Lorentz invariance* condition of the *norm* of the 4-vector. (The difference between this relation and Eq. (52), pertaining to the Euclidian geometry, is the reason why the Minkowski space is called *pseudo-Euclidian*.) It is also straightforward to use Eqs. (51) and (54) to check that an evident generalization of the norm, the *scalar product* of two arbitrary 4-vectors,

Scalar
4-product

$$A_0 B_0 - \sum_{j=1}^3 A_j B_j, \quad (9.56)$$

is also Lorentz-invariant.

Now consider the 4-vector corresponding to a infinitesimal *interval* between two close world events:

$$\{dx_0, dx_1, dx_2, dx_3\} = \{cdt, d\mathbf{r}\}; \quad (9.57)$$

its norm,

Interval
between
two close
events

$$(ds)^2 \equiv dx_0^2 - \sum_{j=1}^3 dx_j^2 = c^2 (dt)^2 - (d\mathbf{r})^2, \quad (9.58)$$

²³ After H. Minkowski who was first to recast (in 1907) the special relativity relations in a form in which space coordinates and time (or rather ct) are treated on an equal footing.

is of course also Lorentz-invariant. Since the speed of any particle (or signal) cannot be larger than c , for any pair of world events that are in a causal relation with each other, dr cannot be larger than cdt , i.e. such *time-like* interval $(ds)^2$ cannot be negative. The 4D surface separating such intervals from space-like intervals $(ds)^2 < 0$ is called the light cone (Fig. 9).

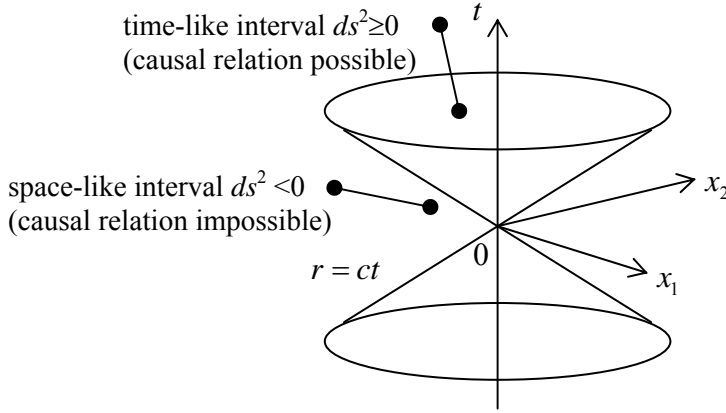


Fig. 9.9. 2+1 dimensional image of the light cone (which is actually 3+1 dimensional).

Now let us assume that these two close world events happen with the same particle that moves with velocity \mathbf{u} . Then in the frame moving with a particle ($\mathbf{v} = \mathbf{u}$), the last term in the right-hand part of Eq. (58) equals zero, so that

$$ds = cd\tau, \quad (9.59)$$

where $d\tau$ is the proper time interval. But according to Eq. (21), this means that we can write

$$d\tau = \frac{dt}{\gamma}, \quad (9.60)$$

where dt is the time interval in an *arbitrary* (besides being inertial) reference frame, while

$$\beta \equiv \frac{\mathbf{u}}{c} \quad \text{and} \quad \gamma \equiv \frac{1}{(1 - \beta^2)^{1/2}} = \frac{1}{(1 - u^2/c^2)^{1/2}} \quad (9.61)$$

are the parameters (17) corresponding to *particle's* velocity (\mathbf{u}) in that frame, so that $ds = cdt/\gamma$.²⁴

Now, let us explore whether a 4-vector can be formed using spatial components of particle's velocity

$$\mathbf{u} = \left\{ \frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt} \right\}. \quad (9.62)$$

Here we have a slight problem: as Eqs. (22) show, these components do not obey the Lorentz transform. However, let us use $d\tau \equiv dt/\gamma$, the proper time interval of the particle, to form the following string:

²⁴ I have opted against using special indices (e.g., β_u , γ_u) to distinguish Eqs. (17) and (61) here and below, in a hope that the suitable velocity (of a reference frame or of a particle) will be always clear from the context.

4-velocity

$$\left\{ \frac{dx_0}{d\tau}, \frac{dx_1}{d\tau}, \frac{dx_2}{d\tau}, \frac{dx_3}{d\tau} \right\} = \gamma \left\{ c, \frac{dx}{dt}, \frac{dy}{dt}, \frac{dz}{dt} \right\} = \gamma \{c, \mathbf{u}\}. \quad (9.63)$$

As follows from comparison of the first form of this expression with Eq. (48), since the time-space vector obeys the Lorentz transform, and τ is Lorentz-invariant, string (63) is a legitimate 4-vector. It is called the *4-velocity* of the particle.

Now we are properly equipped to proceed to dynamics. Let us start with such basic notions of momentum \mathbf{p} and energy \mathcal{E} – so far, for a free particle.²⁵ Perhaps the most elegant way to “derive” (or rather guess²⁶) expressions for \mathbf{p} and \mathcal{E} as functions of particle’s velocity \mathbf{u} is based on analytical mechanics. Due to the conservation of \mathbf{v} , the trajectory of a free particle in the 4D Minkowski space is always a straight line. Hence, from the Hamilton principle of minimum action,²⁷ we may expect its action \mathcal{S} , between points 1 and 2, to be a linear function of the space-time interval (59):

Free
particle’s
action

$$\mathcal{S} = \alpha \int_1^2 ds = \alpha c \int_1^2 d\tau = \alpha c \int_{t_1}^{t_2} \frac{dt}{\gamma}, \quad (9.64)$$

where α is some constant. On the other hand, in analytical mechanics the action is defined as

$$\mathcal{S} = \int_{t_1}^{t_2} \mathcal{L} dt, \quad (9.65)$$

where \mathcal{L} is particle’s Lagrangian function.²⁸ Comparing these two expressions, we get

$$\mathcal{L} = \frac{\alpha c}{\gamma} = \alpha c \left(1 - \frac{u^2}{c^2} \right)^{1/2}. \quad (9.66)$$

In the nonrelativistic limit ($u \ll c$), this function tends to

$$\mathcal{L} \approx \alpha c \left(1 - \frac{u^2}{2c^2} \right) = \alpha c - \frac{\alpha u^2}{2c}. \quad (9.67)$$

In order to correspond to the Newtonian mechanics, the last (velocity-dependent) term should equal $mu^2/2$. From here we find $\alpha = -mc$, so that, finally,

Free
particle’s
Lagrangian
function

$$\mathcal{L} = -mc^2 \left(1 - \frac{u^2}{c^2} \right)^{1/2}. \quad (9.68)$$

²⁵ I am sorry for using, as in Sec. 6.3, for particle’s momentum, the same traditional notation (\mathbf{p}) as had been used for the dipole electric moment. However, since the latter notion will be virtually unused in the balance of the notes, this may hardly lead to confusion.

²⁶ Indeed, such a derivation uses additional assumptions, however natural (such as the Lorentz-invariance of \mathcal{S}), so it can hardly be considered as a real proof of the final results, so that they require experimental confirmation. Fortunately, such confirmations have been numerous – see below.

²⁷ See, e.g., CM Sec. 10.3.

²⁸ See, e.g., CM Sec. 2.1.

Now we can find Cartesian components p_j of particle's momentum, as the generalized momenta corresponding to components r_j ($j = 1, 2, 3$) of the 3D radius-vector \mathbf{r} :²⁹

$$p_j = \frac{\partial \mathcal{L}}{\partial \dot{r}_j} = \frac{\partial \mathcal{L}}{\partial u_j} = -mc^2 \frac{\partial}{\partial u_j} \left(1 - \frac{u_1^2 + u_2^2 + u_3^2}{c^2} \right)^{1/2} = \frac{mu_j}{(1 - u^2/c^2)^{1/2}} = m\gamma u_j. \quad (9.69)$$

Thus for the 3D vector of momentum, we can write the result in the same form as in nonrelativistic mechanics,

$$\mathbf{p} = m\gamma \mathbf{u} = M\mathbf{u}, \quad (9.70) \quad \text{Relativistic momentum}$$

if we introduce the reference-frame-dependent scalar M (called the *relativistic mass*) defined as

$$M \equiv m\gamma = \frac{m}{(1 - u^2/c^2)^{1/2}} \geq m, \quad (9.71) \quad \text{Relativistic mass}$$

m being the nonrelativistic mass of the particle. (It is also called the *rest mass*, because in the reference frame in that the particle rests, Eq. (71) yields $M = m$.)

Next, let us return to analytical mechanics to calculate particle's energy \mathcal{E} (which for a free particle coincides with the Hamiltonian function \mathcal{H}):²⁷

$$\mathcal{E} = \mathcal{H} = \sum_{j=1}^3 p_j u_j - \mathcal{L} = \mathbf{p} \cdot \mathbf{u} - \mathcal{L} = \frac{mu^2}{(1 - u^2/c^2)^{1/2}} + mc^2 \left(1 - \frac{u^2}{c^2} \right) = \frac{mc^2}{(1 - u^2/c^2)^{1/2}}. \quad (9.72)$$

Thus, we have arrived at the most famous of Einstein's formulas (and probably the most famous formula of physics as a whole),

$$\mathcal{E} = m\gamma c^2 = Mc^2, \quad (9.73) \quad \mathcal{E} = Mc^2$$

that expresses the relation between particle's mass and energy.³⁰ In the nonrelativistic limit, it reduces to

$$\mathcal{E} = \frac{mc^2}{(1 - u^2/c^2)^{1/2}} \approx mc^2 \left(1 + \frac{u^2}{2c^2} \right) = mc^2 + \frac{mu^2}{2}, \quad (9.74)$$

the first term mc^2 being called the *rest energy* of a particle.

Now let us consider the following string of 4 scalars:

$$\left\{ \frac{\mathcal{E}}{c}, p_1, p_2, p_3 \right\} = \left\{ \frac{\mathcal{E}}{c}, \mathbf{p} \right\}. \quad (9.75) \quad \text{4-vector of energy-momentum}$$

Using Eqs. (70) and (73) to present this expression as

²⁹ See, e.g., CM Sec. 2.3.

³⁰ Let me hope that the reader understands that all the layman talk about the “mass to energy conversion” is only valid in a very limited sense of the word. While the Einstein relation (73) does allow the conversion of “massive” particles (with $m \neq 0$) into massless particles such as photons, each of the latter particles also has a nonvanishing relativistic mass M , and *simultaneously* the energy \mathcal{E} related to M by Eq. (73).

$$\left\{ \frac{\mathcal{E}}{c}, \mathbf{p} \right\} = m\gamma \{c, \mathbf{u}\}, \quad (9.76)$$

and comparing the result with Eq. (63), we immediately see that, since m is Lorentz-invariant, this string is a legitimate 4-vector of *energy-momentum*. As a result, its norm,

$$\left(\frac{\mathcal{E}}{c} \right)^2 - p^2, \quad (9.77)$$

is Lorentz-invariant, and in particular has to be equal to the norm in the particle rest frame. But in that frame, $p = 0$, and $\mathcal{E} = mc^2$, so that in an arbitrary frame

$$\left(\frac{\mathcal{E}}{c} \right)^2 - p^2 = (mc)^2. \quad (9.78a)$$

This very important relation³¹ between the relativistic energy and momentum (valid for free particles only!) is usually presented in the form³²

$$\mathcal{E}^2 = (mc^2)^2 + (pc)^2. \quad (9.78b)$$

Free
particle's
energy

According to Eq. (70), in the *ultrarelativistic limit* $u \rightarrow c$, p tends to infinity while mc^2 stays constant, so that $pc \gg mc^2$. As follows from Eq. (78), in this limit $\mathcal{E} \approx pc$. Though the above discussion was for particles with finite m , the 4-vector formalism allows us to consider particles with zero rest mass as ultrarelativistic particles for which the above energy-to-moment relation,

$$\mathcal{E} = pc, \quad \text{for } m = 0, \quad (9.79)$$

is exact. Quantum electrodynamics³³ tells us that under certain conditions, electromagnetic field quanta (photons) may be also considered as such *massless* particles, with momentum $\mathbf{p} = \hbar \mathbf{k}$. Plugging (the modulus of) the last relation into Eq. (78), for photon's energy we get $\mathcal{E} = pc = \hbar kc = \hbar \omega$. Please note that according to Eq. (73), the relativistic mass of a photon is finite: $M = \mathcal{E}/c^2 = \hbar \omega/c^2$, so that the term “massless particle” has a limited meaning: $m = 0$. For example, the mass of an optical phonon is of the order of 10^{-36} kg. This is not too much, but still a noticeable (approximately one-millionth) part of the rest mass m_e of an electron.

The fundamental relations (70) and (73) have been repeatedly verified in numerous particle collision experiments, in which the total energy and momentum of a system of particles are conserved – at the same conditions as in the nonrelativistic dynamics. (For momentum, this is the absence of external forces, and for energy, the elasticity of particle interactions – in other words, the absence of alternative channels of energy escape.) Of course, generally, the total energy of the system is conserved, including the potential energy of particle interactions. However, at typical particle collisions, the potential energy

³¹ Please note one more useful relation following from Eqs. (70) and (73): $\mathbf{p} = (\mathcal{E}/c^2)\mathbf{u}$.

³² It may be tempting to interpret this relation as the perpendicular-vector-like addition of the rest energy mc^2 and the “kinetic energy” pc , but from the point of view of the total energy conservation (see below), a better definition of the kinetic energy is $T(u) \equiv \mathcal{E}(u) - \mathcal{E}(0)$.

³³ Briefly reviewed in QM Chapter 9.

vanishes so rapidly with the distance between them that we can use the momentum and energy conservation using Eq. (73).

As an example, let us calculate the minimum energy \mathcal{E}_{\min} of a proton (p_a), necessary for the well-known high-energy reaction that generates a new proton-antiproton pair, $p_a + p_b \rightarrow p + p + p + \bar{p}$, provided that before the collision, proton p_b has been at rest in the lab frame. This minimum evidently corresponds to the vanishing relative velocity of the reaction products, i.e. their motion with virtually the same velocity (\mathbf{u}_{fin}), as seen from the lab frame – see Fig. 10.

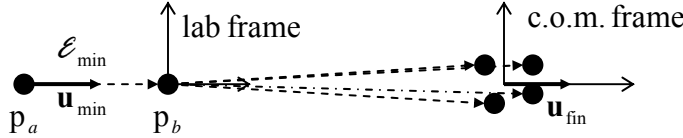


Fig. 9.10. High-energy proton reaction at $\mathcal{E} \approx \mathcal{E}_{\min}$ – schematically.

Due to the momentum conservation, this velocity should have the same direction as the initial velocity (\mathbf{u}_{\min}) of proton p_a . This is why two scalar equations: of for the energy conservation,

$$\frac{mc^2}{(1 - u_{\min}^2 / c^2)^{1/2}} + mc^2 = \frac{4mc^2}{(1 - u_{\text{fin}}^2 / c^2)^{1/2}}, \quad (9.80a)$$

and momentum conservation,

$$\frac{mu}{(1 - u_{\min}^2 / c^2)^{1/2}} + 0 = \frac{4mu_{\text{fin}}}{(1 - u_{\text{fin}}^2 / c^2)^{1/2}}, \quad (9.80b)$$

are sufficient to find both u_{\min} and u_{fin} . After a conceptually simple but rather tedious solution of this system of two nonlinear equations, we get

$$u_{\min} = \frac{4\sqrt{3}}{7}c, \quad u_{\text{fin}} = \frac{\sqrt{3}}{2}c. \quad (9.81)$$

Finally, we can use Eq. (73) to calculate the required energy: $\mathcal{E}_{\min} = 7mc^2$. (Note that of the acceleration energy $6mc^2$, only $2mc^2$ go into the “useful” proton-antiproton pair production.) Proton’s rest mass, $m_p \approx 1.67 \times 10^{-27}$ kg, corresponds to the rest energy $mc^2 \approx 1.502 \times 10^{-10}$ J ≈ 0.938 GeV, so that $\mathcal{E}_{\min} \approx 6.57$ GeV.

The second, more intelligent way to solve the same problem is to use the center-of-mass (*c.o.m.*) reference frame that, in relativity, is defined as the frame in which the total momentum of the system vanishes.³⁴ In this frame, at $\mathcal{E} = \mathcal{E}_{\min}$, the velocity and momenta of all reaction products are equal to zero, while velocities of protons p_a and p_b before the collision are equal and opposite, with some magnitude u' . Hence the energy conservation law becomes

$$\frac{2mc^2}{(1 - u'^2 / c^2)^{1/2}} = 4mc^2, \quad (9.82)$$

³⁴ Note that according to this definition, the c.o.m.’s radius-vector is $\mathbf{R} = \sum_k M_k \mathbf{r}_k / \sum_k M_k = \sum_k \gamma_k m_k \mathbf{r}_k / \sum_k \gamma_k m_k$, generally different from the well-known expression $\sum_k m_k \mathbf{r}_k / \sum_k m_k$ of the nonrelativistic mechanics.

readily giving $u' = (\sqrt{3}/2) c$. This is of course the same result as Eq. (81) gives for u_{fin} . Now we can use the fact that the velocity of proton p_b in the c.o.m. frame is $(-u')$, and hence the speed of proton p_a is $(+u')$. Hence we may find the lab-frame speed of proton p_a using the velocity transform formula (25):

$$u_{\text{min}} = \frac{2u'}{1 + u'^2 / c^2}. \quad (9.83)$$

This relation gives the same result as the first method, $u_{\text{min}} = (4\sqrt{3}/7)c$, but in a much simpler way.

9.4. More on 4-vectors and 4-tensors

This is a good moment to describe a formalism that will allow us, in particular, to solve the same proton collision problem in one more (and arguably, the most elegant) way. More importantly, this formalism will be virtually necessary for the description of the Lorentz transform of the electromagnetic field, and its interaction with relativistic particles – otherwise the formulas would be too cumbersome. Let us call the 4-vectors we have used before,

Contravariant
and
covariant
4-vectors

$$A^\alpha \equiv \{A_0, \mathbf{A}\}, \quad (9.84)$$

contravariant, and denote them with the top index, and introduce also *covariant* vectors,

$$A_\alpha \equiv \{A_0, -\mathbf{A}\}, \quad (9.85)$$

marked by the lower index. Now if we form a scalar product of these vectors using the *standard* (3D-like) rule, just as a sum of the products of the corresponding components, we immediately get

$$A_\alpha A^\alpha \equiv A^\alpha A_\alpha \equiv A_0^2 - A^2. \quad (9.86)$$

Here and below the sign of sum of four components of the product has been dropped.³⁵

The scalar product (86) is just the norm of the 4-vector in our former definition, and as we already know, is Lorentz-invariant. Moreover, the scalar product of two different vectors (also a Lorentz invariant), may be written in any of two similar forms:³⁶

Scalar
product
forms

$$A_0 B_0 - \mathbf{A} \cdot \mathbf{B} \equiv A_\alpha B^\alpha = A^\alpha B_\alpha; \quad (9.87)$$

again, the only caveat is to take one vector in the covariant, and another in the contravariant form.

Now let us return to our sample problem (Fig. 10). Since all components (\mathcal{E}/c and \mathbf{p}) of the total 4-momentum of our system are conserved at the collision, its norm is conserved as well:

$$(p_a + p_b)_\alpha (p_a + p_b)^\alpha = (4p)_\alpha (4p)^\alpha. \quad (9.88)$$

Since now the vector product is the usual math construct, we know that the parentheses in the left-hand part of this equation may be multiplied as usual. We may also swap the operands and move constant factors around as convenient. As a result, we get

³⁵ This compact notation may take some time to be accustomed to, but can hardly lead to any confusion, due to the following rule: the summation is implied always (and only) when an index is repeated twice, once on the top and another at the bottom. In these notes, this shorthand notation will be used only for 4-vectors, but not for the usual (spatial) vectors.

³⁶ Note also that, by definition, for any two 4-vectors, $A_\alpha B^\alpha = B^\alpha A_\alpha$.

$$(p_a)_\alpha (p_a)^\alpha + (p_b)_\alpha (p_b)^\alpha + 2(p_a)_\alpha (p_b)^\alpha = 16 p_a p^\alpha. \quad (9.89)$$

Thanks to the Lorentz-invariance of each of the terms, we can calculate each of them in the reference frame we like. For the first two terms in left-hand part, as well as for the right-hand part term, it is beneficial to use the frames in which that particular proton is at rest; as a result, each of the left-hand part terms equals $(mc)^2$, while the right-hand part equals $16(mc)^2$. On the contrary, the last term of the left-hand part is better evaluated in the lab frame, because in that frame the three spatial components of the 4-momentum p_b vanish, and the scalar product is the just the product of scalars \mathcal{E}/c for protons a and b . For the latter proton this is just mc , so that we get a simple equation,

$$(mc)^2 + (mc)^2 + 2 \frac{\mathcal{E}_{\min}}{c} mc = 16(mc)^2, \quad (9.90)$$

immediately giving the final result: $\mathcal{E}_{\min} = 7 mc^2$ we have already had.

Let me hope that this example was a convincing demonstration of the convenience of presenting 4-vectors in the contravariant (84) and covariant (85) forms,³⁷ with Lorentz-invariant norms (86). To be useful for more complex tasks, the formalism should be developed a little bit further. In particular, it is crucial to know how do the 4-vectors change under the Lorentz transform. For contravariant vectors, we already know the answer (54), but let us rewrite it in the new notation:

$$A^\alpha = L^\alpha_\beta A'^\beta. \quad (9.91)$$

Lorentz
transform of
contravariant
vectors

where L^α_β is the *mixed Lorentz tensor* (51):³⁸

$$L^\alpha_\beta = \begin{pmatrix} \gamma & \beta\gamma & 0 & 0 \\ \beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad (9.92)$$

Mixed
Lorentz
tensor

Note that though the position of indices α and β in the Lorentz tensor notation is not crucial, because it is symmetric, it is convenient to place them using the general *index balance rule*: the difference of the numbers of the upper and lower indices should be the same in both parts of any 4-vector/tensor equality, with the top index in the denominator of a fraction counted as a bottom index in the nominator, and vice versa. (Check yourself that all our formulas above do satisfy this rule.)

In order to rewrite Eq. (91) in a more general form that would not depend on the particular orientation of the coordinate axes (Fig. 1), let us use the contravariant and covariant forms of the 4-vector of the time-space interval (57),

³⁷ These forms are 4-vector extensions of the notions of contravariance and covariance introduced in the 1850s by J. Sylvester for the description of 3D vector change at transfer between different reference frames - e.g., axes rotation – cf. Fig. 3. In this case, the contravariance or covariance of a vector is uniquely determined by its nature: if Cartesian coordinates of a vector (such as the nonrelativistic velocity $\mathbf{v} = d\mathbf{r}/dt$) are transformed similarly to the radius-vector \mathbf{r} , it is called contravariant, while the vectors (such as $\partial f/\partial \mathbf{r} \equiv \nabla f$) that require the reciprocal transform, are called covariant. In the Minkowski space, both forms may be used for any 4-vector.

³⁸ Just as 4-vectors, 4-tensors with two top indices are called contravariant, and those with two bottom indices, covariant. Tensors with one top and one bottom index are called mixed.

$$dx^\alpha = \{cdt, d\mathbf{r}\}, \quad dx_\alpha = \{cdt, -d\mathbf{r}\}; \quad (9.93)$$

then its norm (58) may be presented as³⁹

$$(ds)^2 \equiv (cdt)^2 - (dr)^2 = dx^\alpha dx_\alpha = dx_\alpha dx^\alpha. \quad (9.94)$$

Applying Eq. (91) to the contravariant form of vector (93), we get

$$dx^\alpha = L^\alpha_\beta dx'^\beta. \quad (9.95)$$

But, with our new shorthand notation, we can also write the usual rule of differentiation of each component x^α , considering it as a (in our case, linear) function of 4 arguments x'^β , as follows:

$$dx^\alpha = \frac{\partial x^\alpha}{\partial x'^\beta} dx'^\beta. \quad (9.96)$$

Comparing Eqs. (95) and (96), we can rewrite the general Lorentz transform rule (92) in the new form,

$$A^\alpha = \frac{\partial x^\alpha}{\partial x'^\beta} A'^\beta. \quad (9.97a)$$

General
form
of Lorentz
transform

which evidently does not depend on the coordinate axes orientation.

It is straightforward to verify that the reciprocal transform may be presented as

$$A'^\alpha = \frac{\partial x'^\alpha}{\partial x^\beta} A^\beta. \quad (9.97b)$$

Reciprocal
Lorentz
transform

However, the reciprocal transform differs from the direct one only by the sign of the relative velocity of the frames, so that the transform is given by the inverse matrix $\partial x'^\alpha / \partial x^\beta$; for the coordinate choice shown in Fig. 1, the matrix is

$$\frac{\partial x'^\alpha}{\partial x^\beta} = \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad (9.98)$$

³⁹ Another way to write this relation is $(ds)^2 = g_{\alpha\beta} dx^\alpha dx^\beta = g^{\alpha\beta} dx_\alpha dx_\beta$, where double summation over indices α and β is implied, and g is the so-called *metric tensor*,

$$g^{\alpha\beta} \equiv g_{\alpha\beta} \equiv \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix},$$

that may be used, in particular, to transfer a covariant vector into the corresponding contravariant one and back: $A^\alpha = g^{\alpha\beta} A_\beta$, $A_\alpha = g_{\alpha\beta} A^\beta$. The metric tensor plays a key role in general relativity, in which it is affected by gravity – “curved” by particle masses.

Since, according to Eqs. (84)-(85), covariant 4-vectors differ from the contravariant ones by the sign of the spatial components, their direct transform is given by matrix (98). Hence their direct and reciprocal transforms may be represented, respectively, as

$$A_\alpha = \frac{\partial x'^\beta}{\partial x^\alpha} A'_\beta, \quad A'_\alpha = \frac{\partial x^\beta}{\partial x'^\alpha} A_\beta, \quad (9.99)$$

Lorentz
transform of
covariant
vectors

evidently satisfying the index balance rule. (Note that primed quantities are now multiplied, rather than divided as in the contravariant case.) As a sanity check, let us apply this formalism to the scalar product $A_\alpha A^\alpha$. As Eq. (96) shows, the implicit summation notation allows us to multiply and divide any equality by the same partial differential of a coordinate, so that we can write:

$$A_\alpha A^\alpha = \frac{\partial x'^\beta}{\partial x^\alpha} \frac{\partial x^\alpha}{\partial x'^\gamma} A'_\beta A'^\gamma = \frac{\partial x'^\beta}{\partial x'^\gamma} A'_\beta A'^\gamma = \delta_{\beta\gamma} A'_\beta A'^\gamma = A'_\gamma A'^\gamma, \quad (9.100)$$

i.e. the scalar product $A_\alpha A^\alpha$ (as well as $A^\alpha A_\alpha$) is Lorentz-invariant, as it should be.

Now, let us consider the 4-vectors of derivatives. Here we should be very careful. Consider, for example, the following vector operator

$$\frac{\partial}{\partial x^\alpha} \equiv \left\{ \frac{\partial}{\partial(ct)}, \nabla \right\}, \quad (9.101)$$

As was discussed above, the operator is not changed by its multiplication and division by another differential, e.g., $\partial x'^\beta$ (with the corresponding implied summation over β), so that

$$\frac{\partial}{\partial x^\alpha} = \frac{\partial x'^\beta}{\partial x^\alpha} \frac{\partial}{\partial x'^\beta}. \quad (9.102)$$

But, according to the first of Eqs. (99), this is exactly how the covariant vectors are Lorentz-transformed! Hence, we have to consider the derivative over a *contravariant* space-time interval as a *covariant* 4-vector, and vice versa.⁴⁰ (This result might be also expected from the index balance rule.) In particular, this means that the scalar product

$$\frac{\partial}{\partial x^\alpha} A^\alpha \equiv \frac{\partial A_0}{\partial(ct)} + \nabla \cdot \mathbf{A} \quad (9.103)$$

should be Lorentz-invariant for any legitimate 4-vector. A convenient shorthand for the covariant derivative, which complies with the index balance rule, is

$$\frac{\partial}{\partial x^\alpha} \equiv \partial_\alpha, \quad (9.104)$$

so that the invariant scalar product may be written just as $\partial_\alpha A^\alpha$. A similar definition of the contravariant derivative,

Shorthand
for
4-derivatives

$$\partial^\alpha \equiv \frac{\partial}{\partial x_\alpha} = \left\{ \frac{\partial}{\partial(ct)}, -\nabla \right\}, \quad (9.105)$$

⁴⁰ As was mentioned above, this is also a property of the “usual” transform of 3D vectors.

allows us to write the Lorentz-invariant scalar product (103) in any of two forms:

$$\frac{\partial A_0}{\partial(ct)} + \nabla \cdot \mathbf{A} = \partial^\alpha A_\alpha = \partial_\alpha A^\alpha. \quad (9.106)$$

Finally, let us see how does the general Lorentz transform changes 4-tensors. A second-rank 4×4 matrix is a legitimate 4-tensor if both 4-vectors it relates obey the Lorentz transform. For example, if two legitimate 4-vectors are related as

$$A^\alpha = T^{\alpha\beta} B_\beta, \quad (9.107)$$

we should require that

$$A'^\alpha = T'^{\alpha\beta} B'_\beta, \quad (9.108)$$

where A^α and A'^α are related by Eqs. (97), while B_β and B'_β by Eqs. (99). This requirement immediately yields

$$T^{\alpha\beta} = \frac{\partial x^\alpha}{\partial x'^\gamma} \frac{\partial x^\beta}{\partial x'^\delta} T'^{\gamma\delta}, \quad T'^{\alpha\beta} = \frac{\partial x'^\alpha}{\partial x^\gamma} \frac{\partial x'^\beta}{\partial x^\delta} T^{\gamma\delta}, \quad (9.109)$$

with the implied summation over two indices, γ and δ . The rules for covariant and mixed tensors are similar.⁴¹

9.5. Maxwell equations in the 4-form

This 4-vector formalism background is already sufficient to analyze the Lorentz transform of the electromagnetic field. Just to warm up, let us consider the continuity equation (4.5),

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{j} = 0, \quad (9.110)$$

which expresses the electric charge conservation, and, as we already know, is compatible with the Maxwell equations. If we now define the contravariant and covariant 4-vectors of electric current as

$$j^\alpha = \{\rho c, \mathbf{j}\}, \quad j_\alpha = \{\rho c, -\mathbf{j}\}, \quad (9.111)$$

then Eq. (110) may be presented in the form

$$\partial^\alpha j_\alpha = \partial_\alpha j^\alpha = 0, \quad (9.112)$$

showing that the continuity equation is *form-invariant*⁴² with respect to the Lorentz transform.

Of course, such equation *form* invariance does not mean that all component *values* of the 4-vectors participating in the equation are the same in both frames! For example, let us have some static charge density ρ in frame 0; then Eq. (97b), applied to the contravariant form of 4-vector (111), reads

⁴¹ It is straightforward to check that transfer between the contravariant and covariant forms of the same tensor may be readily achieved using the same metric tensor g : $T_{\alpha\beta} = g_{\alpha\gamma} T^{\gamma\delta} g_{\delta\beta}$, $T^{\alpha\beta} = g^{\alpha\gamma} T_{\gamma\delta} g^{\delta\beta}$.

⁴² Note that in some texts, the equations preserving their form at a transform are called “covariant”, creating a possibility for confusion with covariant vectors and tensors. On the other hand, calling such *equations* “invariant” does not distinguish them properly from invariant *quantities*, such as scalar products of 4-vectors.

$$j'^{\alpha} = \frac{\partial x'^{\alpha}}{\partial x^{\beta}} j^{\beta}, \quad j^{\beta} = \{\rho c, 0, 0, 0\}. \quad (9.113)$$

Using the explicit form (92) of the reciprocal Lorentz matrix for the coordinate choice shown in Fig. 1, we see that this relation yields

$$\rho' = \gamma \rho, \quad j'_x = -\gamma \beta \rho c = -\gamma v \rho, \quad j'_y = j'_z = 0. \quad (9.114)$$

Since the charge velocity, as observed from frame $0'$, is $(-\mathbf{v})$, the nonrelativistic result would be $\mathbf{j} = -\mathbf{v}\rho$. The additional γ factor in the relativistic results for both charge density and current is caused by the length contraction: $dx' = dx/\gamma$, so that in order to keep the total charge $dQ = \rho d^3r = \rho dx dy dz$ inside the elementary volume $d^3r = dx dy dz$ intact, ρ (and hence j_x) should increase proportionally.

Next, in the end of Chapter 6 we have seen that Maxwell equations for potentials ϕ and \mathbf{A} may be presented in a similar form (6.109), under the Lorenz (again, not “Lorentz” please!) gauge condition (6.108). For the free space, this condition takes the form

$$\nabla \cdot \mathbf{A} + \frac{1}{c^2} \frac{\partial \phi}{\partial t} = 0. \quad (9.115)$$

This expression gives us a hint how to form the 4-vector of potentials:⁴³

$$A^{\alpha} = \left\{ \frac{\phi}{c}, \mathbf{A} \right\}, \quad A_{\alpha} = \left\{ \frac{\phi}{c}, -\mathbf{A} \right\}; \quad (9.116) \quad \text{4-vector of potentials}$$

indeed, these vectors satisfy Eq. (115) in its 4-vector form:

$$\partial^{\alpha} A_{\alpha} = \partial_{\alpha} A^{\alpha} = 0. \quad (9.117) \quad \text{Lorenz gauge in 4-form}$$

Since this scalar product is Lorentz-invariant, and derivatives (104)-(105) are legitimate 4-vectors, this implies that 4-vector (116) is also legitimate, i.e. obeys the Lorentz transform formulas (97), (99). A more convincing evidence of this fact may be obtained from Maxwell equations (6.109) for the potentials. In free space, they may be rewritten as

$$\left[\frac{\partial^2}{\partial (ct)^2} - \nabla^2 \right] \frac{\phi}{c} = \frac{(\rho c)}{\epsilon_0 c^2} \equiv \mu_0 (\rho c), \quad \left[\frac{\partial^2}{\partial (ct)^2} - \nabla^2 \right] \mathbf{A} = \mu_0 \mathbf{j}. \quad (9.118)$$

Using definition (116), these equations may be merged to one:⁴⁴

$$\square A^{\alpha} = \mu_0 j^{\alpha}, \quad (9.119) \quad \text{Maxwell equations for 4-potentials}$$

where \square is the *d'Alembert operator*⁴⁵ that may be presented as either of two scalar products,

$$\square \equiv \frac{\partial^2}{\partial (ct)^2} - \nabla^2 = \partial^{\beta} \partial_{\beta} = \partial_{\beta} \partial^{\beta}. \quad (9.120) \quad \text{D'Alembert operator}$$

⁴³ In the Gaussian units, the scalar potential should not be divided by c .

⁴⁴ In the Gaussian units, coefficient μ_0 in the right-hand part of Eq. (119) should be replaced, as usual, with $4\pi/c$.

⁴⁵ Named after J.-B. d'Alembert (1717-1783). Note that in older textbooks, notation \square^2 may be met for this operator.

and hence is Lorentz-invariant. Because of that, and the fact that the Lorentz transform changes both 4-vectors A^α and j^α in a similar way, Eq. (119) does not depend on the reference frame choice. Thus we have arrived at a key point of this chapter: we see that Maxwell equations are indeed form-invariant with respect to the Lorentz transform. As a by-product, the 4-vector form (119) of these equations (for potentials) is extremely simple – and beautiful.

However, as we have seen in Chapter 7, for many applications the Maxwell equations for field vectors are more convenient; so let us present them in the 4-form. For that, we may express the Cartesian components of the usual (3D) field vector vectors

$$\mathbf{E} = -\nabla\phi - \frac{\partial\mathbf{A}}{\partial t}, \quad \mathbf{B} = \nabla \times \mathbf{A}, \quad (9.121)$$

via those of the potential 4-vector A^α . For example,

$$E_x = -\frac{\partial\phi}{\partial x} - \frac{\partial A_x}{\partial t} = -c \left(\frac{\partial\phi}{\partial x} \frac{1}{c} + \frac{\partial A_x}{\partial(ct)} \right) \equiv -c(\partial^0 A^1 - \partial^1 A^0), \quad (9.122)$$

$$B_x = \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \equiv -(\partial^2 A^3 - \partial^3 A^2). \quad (9.123)$$

Completing similar calculations for other field components, we find that the following asymmetric, contravariant *field-strength tensor*,

$$F^{\alpha\beta} \equiv \partial^\alpha A^\beta - \partial^\beta A^\alpha, \quad (9.124)$$

may be expressed via the field components as follows:⁴⁶

Field-
strength
tensors

$$F^{\alpha\beta} = \begin{pmatrix} 0 & -E_x/c & -E_y/c & -E_z/c \\ E_x/c & 0 & -B_z & B_y \\ E_y/c & B_z & 0 & -B_x \\ E_z/c & -B_y & B_x & 0 \end{pmatrix}, \quad (9.125a)$$

so that the covariant form of the tensor is

$$F_{\alpha\beta} \equiv g_{\alpha\gamma} F^{\gamma\delta} g_{\delta\beta} = \begin{pmatrix} 0 & E_x/c & E_y/c & E_z/c \\ -E_x/c & 0 & -B_z & B_y \\ -E_y/c & B_z & 0 & -B_x \\ -E_z/c & -B_y & B_x & 0 \end{pmatrix}. \quad (9.125b)$$

If this expression looks a bit too bulky, note that as a reward, the pair of *inhomogeneous* Maxwell equations, i.e. the two first equations of the system (6.93), which in free space ($\mathbf{D} = \epsilon_0 \mathbf{E}$, $\mathbf{B} = \mu_0 \mathbf{H}$) may be rewritten as

$$\nabla \cdot \frac{\mathbf{E}}{c} = \mu_0 c \rho, \quad \nabla \times \mathbf{B} - \frac{\partial}{\partial(ct)} \frac{\mathbf{E}}{c} = \mu_0 \mathbf{j}, \quad (9.126)$$

⁴⁶ In Gaussian units, this formula, as well as Eq. (131) for $G^{\alpha\beta}$, does not have factors c in all the denominators.

may be now rewritten in a very simple (and manifestly form-invariant) way,

$$\partial_\alpha F^{\alpha\beta} = \mu_0 j^\beta, \quad (9.127)$$

First
pair of
Maxwell
equations
for tensor F

which is comparable with Eq. (119) in its beauty and simplicity. Somewhat counter-intuitively, the pair of *homogeneous* Maxwell equations,

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0, \quad \nabla \cdot \mathbf{B} = 0, \quad (9.128)$$

look, in the 4-vector notation, a bit more complicated:⁴⁷

$$\partial_\alpha F_{\beta\gamma} + \partial_\beta F_{\gamma\alpha} + \partial_\gamma F_{\alpha\beta} = 0. \quad (9.129)$$

Second
pair of
Maxwell
equations
for tensor F

Note, however, that Eq. (128) may be also represented in a much simpler form,

$$\partial_\alpha G^{\alpha\beta} = 0, \quad (9.130)$$

using the so-called *dual* (and also asymmetric) *tensor*

$$G^{\alpha\beta} = \begin{pmatrix} 0 & B_x & B_y & B_z \\ -B_x & 0 & -E_z/c & E_y/c \\ -B_y & E_z/c & 0 & -E_x/c \\ -B_z & -E_y/c & E_x/c & 0 \end{pmatrix}, \quad (9.131)$$

which may be obtained from $F^{\alpha\beta}$, given by Eq. (125), by the following replacements:

$$\frac{\mathbf{E}}{c} \rightarrow -\mathbf{B}, \quad \mathbf{B} \rightarrow \frac{\mathbf{E}}{c}. \quad (9.132)$$

Besides the proof of the form-invariance of the Maxwell equations, the 4-vector formalism allows us to achieve our initial goal: find out how do the electric and magnetic field component change at the transfer between reference frames. Let us apply to tensor $F^{\alpha\beta}$ the reciprocal Lorentz transform given by the second of Eqs. (109). Generally, it gives, for each field component, a sum of 16 terms, but since (for our choice of coordinates, shown in Fig. 1) there are many zeros in the Lorentz transform matrix, and diagonal components of $F^{\gamma\delta}$ equal zero as well, the calculations are rather doable. Let us calculate, for example, $E'_x \equiv -cF'^{01}$. The only nonvanishing terms in the right-hand part are

$$E'_x = -cF'^{01} = -c \left(\frac{\partial x'^0}{\partial x^1} \frac{\partial x'^1}{\partial x^0} F^{10} + \frac{\partial x'^0}{\partial x^0} \frac{\partial x'^1}{\partial x^1} F^{01} \right) = -c\gamma^2 (\beta^2 - 1) \frac{E_x}{c} = E_x. \quad (9.133)$$

Repeating the calculation for other 5 components of the fields, we get very important relations

$$\begin{aligned} E'_x &= E_x, & B'_x &= B_x, \\ E'_y &= \gamma(E_y - vB_z), & B'_y &= \gamma(B_y + vE_z/c^2), \\ E'_z &= \gamma(E_z + vB_y), & B'_z &= \gamma(B_z - vE_y/c^2), \end{aligned} \quad (9.134)$$

⁴⁷ To be fair, note that just as Eq. (127), Eq. (129) this is also a set of four scalar equations – in the latter case with indices α, β , and γ taking any three *different* values of the set $\{0, 1, 2, 3\}$.

whose more compact “semi-vector” form is

Lorentz
transform of
field
components

$$\begin{aligned} \mathbf{E}'_{\parallel} &= \mathbf{E}_{\parallel}, & \mathbf{B}'_{\parallel} &= \mathbf{B}_{\parallel}, \\ \mathbf{E}'_{\perp} &= \gamma(\mathbf{E} + \mathbf{v} \times \mathbf{B})_{\perp}, & \mathbf{B}'_{\perp} &= \gamma(\mathbf{B} - \mathbf{v} \times \mathbf{E} / c^2)_{\perp}, \end{aligned} \quad (9.135)$$

where indices \parallel and \perp stand, respectively, for the field components parallel and perpendicular to the relative velocity \mathbf{v} of the two reference frames. In the nonrelativistic limit, the Lorentz factor γ tends to 1, and Eqs. (135) acquire an even simpler form

$$\mathbf{E}' \rightarrow \mathbf{E} + \mathbf{v} \times \mathbf{B}, \quad \mathbf{B}' \rightarrow \mathbf{B} - \frac{1}{c^2} \mathbf{v} \times \mathbf{E}. \quad (9.136)$$

Thus we see that the electric and magnetic fields actually transform to each other even in the first order of the v/c ratio. For example, if we fly across the field lines of a uniform, static, purely electric field \mathbf{E} (e.g., the one in a plane capacitor) we will see not only the electric field re-normalization (in the second order of the v/c ratio), but also a nonvanishing dc magnetic field \mathbf{B}' perpendicular to both vector \mathbf{E} and vector \mathbf{v} , the direction of our motion. This is of course what might be expected from the relativity principle: from the point of view of the moving observer (which is as legitimate as that of a stationary observer), the surface charges of capacitor plates, that create field \mathbf{E} , move back creating dc currents (114) which induce the apparent magnetic field. Similarly, motion across a magnetic field creates, from the point of view of the moving observer, an electric field.

This fact is very important philosophically. One can say there is no such thing in Mother Nature as an electric field (or a magnetic field) all by itself. Not only can the electric field induce the magnetic field (and vice versa) in dynamics, but even in an apparently static configuration, what exactly we measure depends on our speed relative to the field sources – hence the very appropriate term for the whole field we are studying: the *electromagnetism*.

Another simple but very important application of Eqs. (134)-(135) is the calculation of the fields created by a charged particle moving in free space by inertia, i.e. along a straight line with constant velocity \mathbf{u} , at the *impact parameter*⁴⁸ (the closest distance) b from the observer. Selecting frame $0'$ to move with the particle in its origin, and frame 0 to reside in the “lab” (in that fields \mathbf{E} and \mathbf{B} are measured), we can take $\mathbf{v} = \mathbf{u}$. In this case fields \mathbf{E}' and \mathbf{B}' may be calculated from, respectively, electro- and magnetostatics, because in frame $0'$ the particle does not move:

$$\mathbf{E}' = \frac{q}{4\pi\epsilon_0} \frac{\mathbf{r}'}{r'^3}, \quad \mathbf{B}' = 0. \quad (9.137)$$

Selecting the coordinate axes so that at the measurement point $x = 0, y = b, z = 0$ (Fig. 11a), we may write $x' = -ut', y' = b, z' = 0$, so that $r' = (u^2 t'^2 + b^2)^{1/2}$, and the field components are as follows:

$$E'_x = -\frac{q}{4\pi\epsilon_0} \frac{ut'}{(u^2 t'^2 + b^2)^{3/2}}, \quad E'_y = \frac{q}{4\pi\epsilon_0} \frac{b}{(u^2 t'^2 + b^2)^{3/2}}, \quad E'_z = 0, \quad B'_x = B'_y = B'_z = 0. \quad (9.138)$$

Now using the last of Eq. (19b), with $x = 0$, for the time transform, and the equations reciprocal to Eqs. (134) for the field transform (it is evident that they are similar to the direct transform with v replaced with $-v = -u$), in the lab frame we get

⁴⁸ This term is very popular in the of particle scattering – see, e.g., CM Sec. 3.7.

$$E_x = E'_x = -\frac{q}{4\pi\epsilon_0} \frac{u\gamma t}{(u^2\gamma^2 t^2 + b^2)^{3/2}}, \quad E_y = \gamma E'_y = \frac{q}{4\pi\epsilon_0} \frac{\gamma b}{(u^2\gamma^2 t^2 + b^2)^{3/2}}, \quad E_z = 0, \quad (9.139)$$

$$B_x = 0, \quad B_y = 0, \quad B_z = \frac{\gamma u}{c^2} E'_y = \frac{u}{c^2} \frac{q}{4\pi\epsilon_0} \frac{\gamma b}{(u^2\gamma^2 t^2 + b^2)^{3/2}} = \frac{u}{c^2} E_y. \quad (9.140)$$

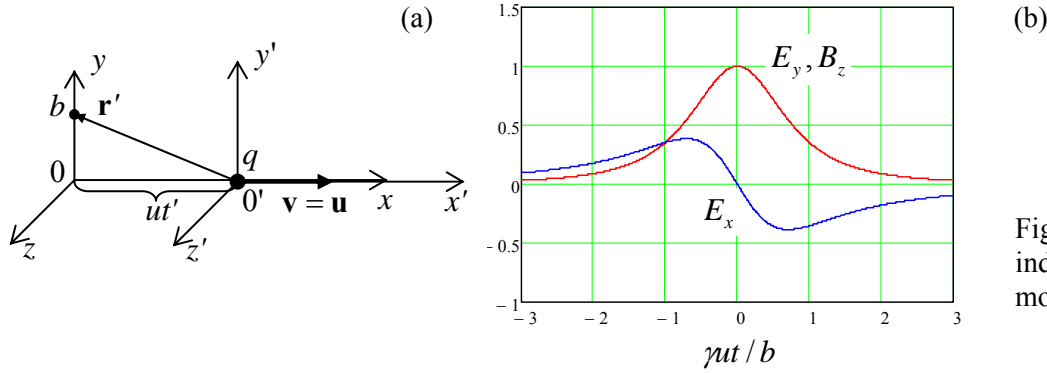


Fig. 9.11. Field pulses induced by a uniformly moving charge.

These results,⁴⁹ plotted in Fig. 11b, reveal two major effects. First, the charge passage by the observer generates not only an electric field pulse, but also a magnetic field pulse. This is natural, because, as was repeatedly discussed in Chapter 5, charge motion is essentially an electric current.⁵⁰ Second, Eqs. (139)-(140) show that the pulse duration scale is

$$\Delta t = \frac{b}{\gamma u} = \frac{b}{u} \left(1 - \frac{u^2}{c^2}\right)^{1/2}, \quad (9.141)$$

i.e. shrinks to zero as the charge velocity u approaches the speed of light. This is of course a direct corollary of the relativistic length contraction: in the frame $0'$ moving with the charge, the longitudinal spread of its electric field at distance b from the motion line is of the order of $\Delta x' = b$. When observed from the lab frame 0 , this interval, in accordance with Eq. (20), shrinks to $\Delta x = \Delta x'/\gamma = b/\gamma$, and so does the pulse duration scale $\Delta t = \Delta x/u = b/\gamma u$.

9.6. Relativistic particles in electric and magnetic fields

Now let us analyze dynamics of charged particles in electric and magnetic fields. Inspired by “our” success of forming the 4-vector (75) of energy-momentum,

$$p^\alpha = \left\{ \frac{\mathcal{E}}{c}, \mathbf{p} \right\} = \gamma \{ mc, \mathbf{p} \} = m \frac{dx^\alpha}{d\tau} \equiv m u^\alpha, \quad (9.142)$$

where u^α is the contravariant form of the 4-velocity (63) of the particle,

⁴⁹ In the next chapter, we will re-derive them in a different way.

⁵⁰ It is straightforward to use Eq. (140) and the linear superposition principle to calculate, for example, the magnetic field of a string of charges moving along the same line, and separated by equal distances $\Delta x = a$ (so that the average current, as measured in frame 0 , is qu/a), and to show that the time-average of the magnetic field is given by Eq. (5.20) of magnetostatics, with b instead of ρ .

$$u^\alpha \equiv \frac{dx^\alpha}{d\tau}, \quad u_\alpha \equiv \frac{dx_\alpha}{d\tau}, \quad (9.143)$$

we may notice that the nonrelativistic equation of motion, resulting from the Lorentz-force formula (5.10) for the three spatial components of p^α , at charged particle's motion in electromagnetic field,

Charged
particle's
dynamics

$$\frac{d\mathbf{p}}{dt} = q(\mathbf{E} + \mathbf{u} \times \mathbf{B}), \quad (9.144)$$

is fully consistent with the following 4-vector equality (which is evidently form-invariant):

Particle's
dynamics
in 4-form

$$\frac{dp^\alpha}{d\tau} = qF^{\alpha\beta}u_\beta. \quad (9.145)$$

For example, the $\alpha = 1$ component of this equation reads

$$\frac{dp^1}{d\tau} = qF^{1\beta}u_\beta = q\left[\frac{E_x}{c}\gamma c + 0 \cdot (-\gamma u_x) + (-B_z)(-\gamma u_y) + B_y(-\gamma u_z)\right] = q\gamma[\mathbf{E} + \mathbf{u} \times \mathbf{B}]_x, \quad (9.146)$$

and similarly for two other spatial components ($\alpha = 2$ and $\alpha = 3$). We see that these expressions differ from the Newton law (144) by the extra factor γ . However, plugging into Eq. (146) the definition of the proper time interval, $d\tau = dt/\gamma$, and canceling γ in both parts, we recover Eq. (144) exactly – for *any* velocity of the particle! The only caveat is that if u is comparable with c , \mathbf{p} in Eq. (144) has to be understood as the relativistic momentum (70) proportional to the velocity-dependent mass $M = \gamma m \geq m$ rather than to the rest mass m .

The only remaining task is to examine the meaning of the 0th component of Eq. (145). Let us spell it out:

$$\frac{dp^0}{d\tau} = qF^{0\beta}u_\beta = q\left[0 \cdot \gamma c + \left(-\frac{E_x}{c}\right)(-\gamma u_x) + \left(-\frac{E_y}{c}\right)(-\gamma u_y) + \left(-\frac{E_z}{c}\right)(-\gamma u_z)\right] = q\gamma \frac{\mathbf{E} \cdot \mathbf{u}}{c}. \quad (9.147)$$

Recalling that $p^0 = \mathcal{E}/c$, and using $d\tau = dt/\gamma$ again, we see that Eq. (147) looks exactly as the nonrelativistic relation for the kinetic energy change,⁵¹

Particle's
energy
evolution

$$\frac{d\mathcal{E}}{dt} = q\mathbf{E} \cdot \mathbf{u}, \quad (9.148)$$

besides that in the relativistic case the energy has to be taken in the general form (73).

No question, the 4-component equation (145) of relativistic dynamics is beautiful in its simplicity. However, for the solution of particular problems, Eqs. (144) and (148) are frequently preferable. As an illustration of this point, let us now use these equations to explore the relativistic effects at charged particle motion in uniform, time-independent electric and magnetic fields. In doing that, we will, for the time being, neglect the contributions into the field by the particle itself.⁵²

⁵¹ See, e.g., CM Eq. (1.20) with $d\mathbf{p}/dt = \mathbf{F} = q\mathbf{E}$. (As a reminder, the magnetic field cannot affect particle's energy, because the magnetic component of the Lorentz force is perpendicular to its velocity.)

⁵² As was emphasized earlier in this course, in statics this contribution has to be ignored. In dynamics, this is generally not true; these *self-action effects* will be discussed in Sec. 10.6.

(i) Uniform magnetic field. Let the magnetic field be constant and uniform in the “lab” reference frame 0. Then in this frame, Eqs. (144) and (148) yield

$$\frac{d\mathbf{p}}{dt} = q\mathbf{u} \times \mathbf{B}, \quad \frac{d\mathcal{E}}{dt} = 0. \quad (9.149)$$

From the second equation, $\mathcal{E} = \text{const}$, we get $u = \text{const}$, $\beta \equiv u/c = \text{const}$, $\gamma \equiv (1 - \beta^2)^{-1/2} = \text{const}$, and $M \equiv \gamma m = \text{const}$, so that the first of Eqs. (149) may be rewritten as

$$\frac{d\mathbf{u}}{dt} = \mathbf{u} \times \boldsymbol{\omega}_c, \quad (9.150)$$

where $\boldsymbol{\omega}_c$ is the vector directed along the magnetic field \mathbf{B} , with the magnitude equal to the *cyclotron frequency* (sometimes called “gyrofrequency”)

$$\omega_c \equiv \frac{qB}{M} = \frac{qB}{\gamma m} = \frac{qc^2 B}{\mathcal{E}}. \quad (9.151) \quad \text{Cyclotron frequency}$$

If particle’s initial velocity \mathbf{u}_0 is perpendicular to the magnetic field, Eq. (150) describes its circular motion, with a constant speed $u = u_0$, in a plane perpendicular to \mathbf{B} , and frequency (151). In the nonrelativistic limit $u \ll c$, when $\gamma \rightarrow 1$, i.e. $M \rightarrow m$, the cyclotron frequency is independent on the speed, but as the kinetic energy is increased to comparable to the rest energy of the particle, the frequency decreases, and in the ultrarelativistic limit,

$$\omega_c \approx qc \frac{B}{p}, \quad \text{at } u \approx c. \quad (9.152)$$

The cyclotron motion radius may be calculated as $R = u/\omega_c$; in the nonrelativistic limit it is proportional to particle’s speed, i.e. to the square root of its kinetic energy. However, in the general case the radius is proportional to particle’s relativistic momentum rather than its speed:

$$R = \frac{u}{\omega_c} = \frac{Mu}{qB} = \frac{m\gamma u}{qB} = \frac{1}{q} \frac{p}{B}, \quad (9.153) \quad \text{Cyclotron radius}$$

so that in the ultrarelativistic limit, when $p \approx \mathcal{E}/c$, R is proportional to the kinetic energy.

This dependence of ω_c and R on energy are the major factors in design of circular accelerators of charged particles. In the simplest of these machines (the *cyclotron*, invented in 1929 by E. Lawrence), frequency ω of the accelerating ac electric field is constant, so that even it is tuned to ω_c of the initially injected particles, the drop of the cyclotron frequency with energy eventually violates this tuning. Due to this reason, the maximum particle speed is limited to just $\sim 0.1 c$ (for protons, corresponding to the kinetic energy of just ~ 15 MeV). This problem may be addressed in several ways. In particular, in *synchrotrons* (such as Fermilab’s Tevatron and CERN’s LHC) the magnetic field is gradually increased in time to compensate the momentum increase ($B \propto p$), so that both R (148) and ω_c (147) stay constant, enabling proton acceleration to energies as high as ~ 7 TeV, i.e. $\sim 2,000 mc^2$.⁵³

⁵³ For more on this topic, I have to refer the interested reader to special literature, for example either S. Lee, *Accelerator Physics*, 2nd ed., World Scientific, 2004, or E. Wilson, *An Introduction to Particle Accelerators*, Oxford U. Press, 2001.

Returning to our initial problem, if particle's initial velocity has a component u_{\parallel} along the magnetic field, it is conserved in time, so that the trajectory is a spiral around the magnetic field lines. As Eqs. (149) show, in this case Eq. (150) remains valid, but in Eqs. (151) and (153) the full speed and momentum have to be replaced with magnitudes of their (also time-conserved) components, u_{\perp} and p_{\perp} , normal to \mathbf{B} , while the Lorentz factor γ in those formulas still requires the full speed of the particle.

Finally, in the special case when particle's initial velocity is directed *exactly* along the magnetic field's direction, it continues to move by straight line along vector \mathbf{B} . In this case, the cyclotron frequency (151) remains finite, but does not correspond to any real motion, because $R = 0$.

(ii) Uniform electric field. This problem is (technically) more complex than the previous one, because in the electric field, particle's kinetic energy may change. Directing axis z along the field, from Eq. (144) we get

$$\frac{dp_z}{dt} = qE, \quad \frac{d\mathbf{p}_{\perp}}{dt} = 0. \quad (9.154)$$

If the field does not change in time, the first integration of these equations is trivial,

$$p_z(t) = p_z(0) + qEt, \quad \mathbf{p}_{\perp}(t) = \text{const} = \mathbf{p}_{\perp}(0), \quad (9.155)$$

but the further integration requires care, because the effective mass $M = \gamma m$ of the particle depends on its full speed:

$$u^2 = u_z^2 + u_{\perp}^2, \quad (9.156)$$

making the two motions, along and across the field, mutually dependent.

If the initial velocity is perpendicular to field \mathbf{E} , i.e. if $p_z(0) = 0$, $p_{\perp}(0) = p(0) \equiv p_0$, the easiest way to proceed is to calculate the kinetic energy first:

$$\mathcal{E}^2 = (mc^2)^2 + c^2 p^2(t) = \mathcal{E}_0^2 + c^2 (qEt)^2, \quad \text{where } \mathcal{E}_0 \equiv [(mc^2)^2 + c^2 p_0^2]^{1/2}. \quad (9.157)$$

On the other hand, we can calculate the same energy by integrating Eq. (148),

$$\frac{d\mathcal{E}}{dt} = q\mathbf{E} \cdot \mathbf{u} = qE \frac{dz}{dt}, \quad (9.158)$$

over time, with a simple result:

$$\mathcal{E} = \mathcal{E}_0 + qEz(t), \quad (9.159)$$

where (for the notation simplicity) I took $z(0) = 0$. Requiring Eq. (159) to give the same \mathcal{E}^2 as Eq. (157), we get a quadratic equation for $z(t)$,

$$\mathcal{E}_0^2 + c^2 (qEt)^2 = [\mathcal{E}_0 + qEz(t)]^2, \quad (9.160)$$

whose solution (with the sign before the square root corresponding to $E > 0$, i.e. $z \geq 0$) is

$$z(t) = \frac{\mathcal{E}_0}{qE} \left\{ \left[1 + \left(\frac{cqEt}{\mathcal{E}_0} \right)^2 \right]^{1/2} - 1 \right\}. \quad (9.161)$$

Now let us find particle's trajectory. Selecting axis x so that the initial velocity vector (and hence the velocity vector at any further instant) is within the $[x, z]$ plane, i.e. $y(t) \equiv 0$, we may use Eqs. (155) to calculate trajectory's slope, at its arbitrary point, as

$$\frac{dz}{dx} = \frac{dz/dt}{dx/dt} = \frac{Mu_z}{Mu_x} = \frac{p_z}{p_x} = \frac{qEt}{p_0}. \quad (9.162)$$

Now let us use Eq. (160) to express the nominator of this fraction, qEt , as a function of z :

$$qEt = \frac{1}{c} \left[(\mathcal{E}_0 + qEz)^2 - \mathcal{E}_0^2 \right]^{1/2}. \quad (9.163)$$

Plugging this expression into Eq. (161), we get

$$\frac{dz}{dx} = \frac{1}{cp_0} \left[(\mathcal{E}_0 + qEz)^2 - \mathcal{E}_0^2 \right]^{1/2}. \quad (9.164)$$

This differential equation may be readily integrated, separating variables z and x , and using substitution $\xi \equiv \text{arccosh}(qEz/\mathcal{E}_0 + 1)$. Selecting the origin of axis x at the initial point, so that $x(0) = 0$, we finally get the trajectory:

$$z = \frac{\mathcal{E}_0}{qE} \left(\cosh \frac{qEx}{cp_0} - 1 \right). \quad (9.165)$$

At the initial part of the trajectory, where $qEx \ll cp_0(0)$, this expression may be approximated by the first nonvanishing term of the Taylor series, giving a parabola:

$$z = \frac{\mathcal{E}_0 qE}{2} \left(\frac{x}{cp_0} \right)^2, \quad (9.166)$$

so that if the initial velocity of the particle is much less than c (i.e. $p_0 \approx mu_0$, $\mathcal{E}_0 \approx mc^2$), we get the familiar nonrelativistic formula:

$$z = \frac{qE}{2mu_0^2} x^2 = \frac{a}{2} t^2, \quad a = \frac{F}{m} = \frac{qE}{m}. \quad (9.167)$$

This solution may be readily generalized to the case of an arbitrary direction of particle's initial velocity; this generalization is left for reader's exercise.

(iii) Crossed uniform magnetic and electric fields ($\mathbf{E} \perp \mathbf{B}$). In the view of how bulky the solution of the previous problem (i.e. the particular case of the current problem for $\mathbf{B} = 0$) was, one might think that this problem should be forbiddingly complex for an analytical solution. Counter-intuitively, it is not the case, due to the help from the field transform relations (135). Let us consider two possible cases.

Case I: $E/c < B$. Let us consider an inertial frame moving (relatively the "lab" reference frame 0 in which fields \mathbf{E} and \mathbf{B} are defined) with velocity

$$\mathbf{v} = \frac{\mathbf{E} \times \mathbf{B}}{B^2}, \quad (9.168)$$

whose magnitude $v = c \times (E/c)/B < c$. Selecting the coordinate axes as shown in Fig. 12, so that

$$E_x = 0, \quad E_y = E, \quad E_z = 0; \quad B_x = 0, \quad B_y = 0, \quad B_z = 0, \quad (9.169)$$

we see that the Cartesian components of this velocity are $v_x = v$, $v_y = v_z = 0$.

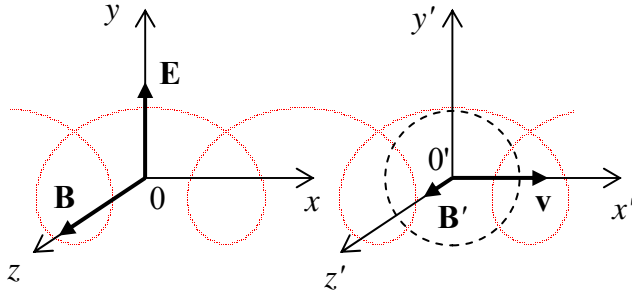


Fig. 9.12. Particle's trajectory in crossed electric and magnetic fields (at $E/c < B$).

Since this choice of coordinates complies with that used to derive Eqs. (134), we can readily use that simple form of the Lorentz transform to calculate field components in the moving reference frame:

$$E'_x = 0, \quad E'_y = \gamma(E - vB) = \gamma\left(E - \frac{E}{B}B\right) = 0, \quad E'_z = 0, \quad (9.170)$$

$$B'_x = 0, \quad B'_y = 0, \quad B'_z = \gamma\left(B - \frac{vE}{c^2}\right) = \gamma B\left(1 - \frac{vE}{Bc^2}\right) = \gamma B\left(1 - \frac{v^2}{c^2}\right) = \frac{B}{\gamma} \leq B, \quad (9.171)$$

where the Lorentz parameter $\gamma \equiv (1 - v^2/c^2)^{-1/2}$ corresponds to velocity (168) rather than that of the particle.

Thus in this special reference frame the particle only sees a (re-normalized) uniform magnetic field $B' \leq B$, parallel to the initial field, i.e. perpendicular to velocity (168). Using the result of the above example (i), we see that in this frame the particle will move along either a circle or a spiral winding about the direction of the magnetic field, with angular speed (151),

$$\omega'_c = \frac{qB'}{\mathcal{E}'/c^2}, \quad (9.172)$$

and radius (148):

$$R' = \frac{p'_\perp}{qB'}. \quad (9.173)$$

Hence in the lab frame, the particle will perform such orbital motion plus a “drift” with constant velocity \mathbf{v} (Fig. 12). As the result, the lab-frame trajectory of the particle (or rather its projection onto the plane perpendicular to the magnetic field) is a *trochoid*-like curve⁵⁴ that, depending on the initial velocity, may be either *prolate* (self-crossing), as in Fig. 12, or *curtate* (stretched so much that it is not self-crossing).

⁵⁴ As a reminder, a trochoid may be described as the trajectory of a point on a rigid disk rolled along a straight line. Its canonical parametric presentation is $x = \Theta + a \cos \Theta$, $y = a \sin \Theta$. (For $a > 1$, the trochoid is *prolate*, if $a < 1$, it is *curtate*, and if $a = 1$, it is called the *cycloid*.) Note, however, that for our problem, the trajectory in the lab frame is exactly trochoidal only in the nonrelativistic limit $v \ll c$ (i.e. $E/c \ll B$), because otherwise the Lorentz contraction in the drift direction squeezes the cyclotron orbit from a circle into an ellipse.

Such looped motion of electrons (in practice, with $v \ll c$) is used, in particular, in *magnetrons* – generators of microwave radiation. In these devices (Fig. 13a), the magnetic field, usually created by specially-shaped permanent magnets, is nearly uniform (in the region of electron motion) and directed along magnetron's axis, while the electric field of magnitude $E \ll cB$, created by the dc voltage applied between the anode and cathode, is virtually radial. As a result, the above simple theory is only approximately valid, and electron trajectories are close to *epicycloids* rather than trochoids. The applied electric field is adjusted so that these trajectories pass close to the gap openings to cylindrical microwave cavities drilled in magnetron's bulk anode (Fig. 13b). The fundamental mode of each cavity is quasi-lumped, with cylindrical walls working mostly as lumped inductances, and gaps as lumped capacitances, with the microwave electric field concentrated in the gap openings. This is why the mode is strongly coupled to the passing electrons, and their interaction creates large positive feedback (equivalent to negative damping) that results in intensive microwave self-oscillations at cavities' eigenfrequency.⁵⁵ The oscillation energy, of course, is taken from the dc-field-accelerated electrons; due to the energy loss each electron gradually moves closer to the anode and finally lands on its surface. The wide use of such generators (in particular, in microwave ovens, which operate in a narrow frequency band around 2.45 GHz, allocated for these devices to avoid their interference with wireless communication systems) is due to their simplicity and high (up to 65%) efficiency.

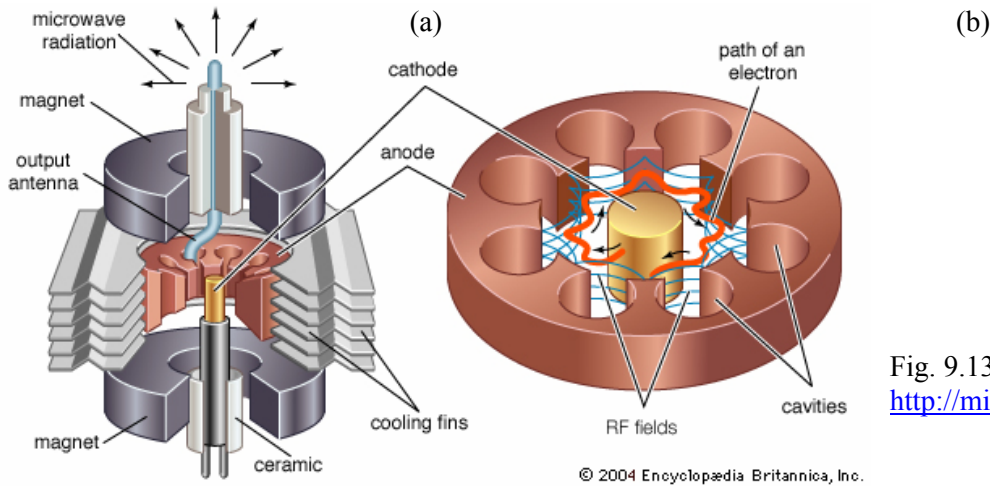


Fig. 9.13. Magnetron. (Adapted from <http://microwavetubes.iwarp.com>.)

© 2004 Encyclopædia Britannica, Inc.

Case II: $E/c > B$. In this case, the speed given by Eq. (168) would be above the speed of light, so let us introduce a reference frame moving with a different velocity,

$$\mathbf{v} = \frac{\mathbf{E} \times \mathbf{B}}{(E/c)^2}, \quad (9.174)$$

whose direction is the same as before (Fig. 12), and magnitude $v = c \times B / (E/c)$ is again below c . A calculation absolutely similar to the one performed above for Case I, yields

$$E'_x = 0, \quad E'_y = \gamma(E - vB) = \gamma E \left(1 - \frac{vB}{E}\right) = \gamma E \left(1 - \frac{v^2}{c^2}\right) = \frac{E}{\gamma} \leq E, \quad E'_z = 0, \quad (9.175)$$

⁵⁵ See, e.g., CM Sec. 4.4.

$$B'_x = 0, \quad B'_y = 0, \quad B'_z = \gamma \left(B - \frac{vE}{c^2} \right) = \gamma \left(B - \frac{EB}{E} \right) = 0. \quad (9.176)$$

so that in the moving frame the particle sees only an electric field $E' \leq E$. According to the solution of our previous problem (ii), the trajectory of the particle in the moving frame is hyperbolic, so that in the lab frame it has an “open”, hyperbolic character as well.

To conclude this section, let me note that if the electric and magnetic fields are non-uniform, the particle motion is much more complex, and in most cases the integration of equations (144), (148) may be carried out only numerically. However, if the field nonuniformity is small, (approximate) analytical methods may be very effective. For example, if the magnetic field has a small *longitudinal* gradient ∇B in a direction perpendicular to vector \mathbf{B} itself, such that

$$\eta \equiv \frac{|\nabla B|}{B} \ll \frac{1}{R}, \quad (9.177)$$

where R is the cyclotron radius (153), then it is straightforward to use Eq. (150) to show⁵⁶ that the cyclotron orbit drifts perpendicular to both \mathbf{B} and ∇B , with speed

$$v_d \approx \frac{\eta}{\omega_c} \left(\frac{1}{2} u_{\perp}^2 + u_{\parallel}^2 \right) \ll u. \quad (9.178)$$

The physics of this drift is rather simple: according to Eq. (153), the instant curvature of the cyclotron orbit is proportional to the local value of the field. Hence if the field is nonuniform, the trajectory bends more on its parts passing through stronger field, thus acquiring a shape close to a curate trochoid.

For engineering and experimental practice, effects of *longitudinal* gradients of magnetic field on charged particle motion are much more important, but let me postpone their discussion until we have got a little bit more analytical tools in the next section.

9.7. Analytical mechanics of charged particles

Equation (145) gives a full description of relativistic particle dynamics in electric and magnetic fields, just as the 2nd Newton law (1) does it in the nonrelativistic limit. However, we know that in the latter case, the Lagrange formalism of analytical mechanics allows an easier solution of many problems.⁵⁷ We can fully expect that to be true in relativistic mechanics as well, so let us expand the analysis of Sec. 3 to particles in the field.

Let recall that for a free particle, our main result was Eq. (68), which may be rewritten as

$$\gamma \mathcal{L} = -mc^2, \quad (9.179)$$

showing that this product is Lorentz-invariant. How can the electromagnetic field affect this relation? In electrostatics, we could write

$$\mathcal{L} = T - U = T - q\phi. \quad (9.180)$$

⁵⁶ See, e.g., Sec. 12.4 in J. D. Jackson, *Classical Electrodynamics*, 3rd ed., Wiley, 1999.

⁵⁷ See, e.g., CM Sec. 2.2 and beyond.

However, in relativity the scalar potential ϕ is just one component of the potential 4-vector (116). The only way to get a Lorentz-invariant contribution to $\gamma\mathcal{L}$ from the full 4-vector, that would be also proportional to the Lorentz force, i.e. to the first power of particle's velocity (to account for the magnetic component of the Lorentz force), is evidently

$$\gamma\mathcal{L} = -mc^2 + \text{const} \times u^\alpha A_\alpha, \quad (9.181)$$

where u^α is the 4-velocity (63). In order to comply with Eq. (180) in electrostatics, the constant factor should be equal to $(-qc)$, so that Eq. (182) becomes

$$\gamma\mathcal{L} = -mc^2 - qu^\alpha A_\alpha, \quad (9.182)$$

and, finally,

$$\mathcal{L} = -\frac{mc^2}{\gamma} - q\phi + q\mathbf{u} \cdot \mathbf{A}, \quad (9.183) \quad \text{Lagrangian function}$$

i.e., in the Cartesian form,

$$\mathcal{L} = -mc^2 \left(1 - \frac{u_x^2 + u_y^2 + u_z^2}{c^2} \right)^{1/2} - q\phi + q(u_x A_x + u_y A_y + u_z A_z). \quad (9.184)$$

Let us see whether this relation (that admittedly was obtained above by an educated guess rather than by a strict derivation) passes a natural sanity check. For the case of unconstrained motion of a particle, we can select its three Cartesian coordinates r_j ($j = 1, 2, 3$) as the generalized coordinates, and linear velocity components u_j as the corresponding generalized velocities. In this case, the Lagrange equations of motion are⁵⁸

$$\frac{d}{dt} \frac{\partial \mathcal{L}}{\partial u_j} - \frac{\partial \mathcal{L}}{\partial r_j} = 0. \quad (9.185)$$

For example, for $r_1 = x$, Eq. (184) yields

$$\frac{\partial \mathcal{L}}{\partial u_x} = \frac{mu_x}{(1 - u^2/c^2)^{1/2}} + qA_x = p_x + qA_x, \quad \frac{\partial \mathcal{L}}{\partial x} = -q \frac{\partial \phi}{\partial x} + q\mathbf{u} \cdot \frac{\partial \mathbf{A}}{\partial x}, \quad (9.186)$$

so that Eq. (185) takes the form

$$\frac{dp_x}{dt} = -q \frac{\partial \phi}{\partial x} + q\mathbf{u} \cdot \frac{\partial \mathbf{A}}{\partial x} - q \frac{dA_x}{dt}. \quad (9.187)$$

In equations of motion, field values have to be taken at the instant position of the particle, so that the last (full) derivative has components due to both the actual field change (at a fixed point of space) and the particle's motion. Such addition is described by the so-called *convective derivative*⁵⁹

$$\frac{d}{dt} = \frac{\partial}{\partial t} + \mathbf{u} \cdot \nabla. \quad (9.188) \quad \text{Convective derivative}$$

⁵⁸ See, e.g., CM Sec. 2.1.

⁵⁹ Alternatively called the “Lagrangian derivative”; for its (rather simple) derivation see, e.g., CM Sec. 8.3.

Spelling out both scalar products, we may group the terms remaining after cancellations as follows:

$$\frac{dp_x}{dt} = q \left[\left(-\frac{\partial \phi}{\partial x} - \frac{\partial A_x}{\partial t} \right) + u_y \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \right) - u_z \left(\frac{\partial A_x}{\partial z} - \frac{\partial A_z}{\partial x} \right) \right]. \quad (9.189)$$

But taking into account relations (121) between the electric and magnetic fields and potentials, this expression is nothing more than

$$\frac{dp_x}{dt} = q(E_x + u_y B_z - u_z B_y) = q(\mathbf{E} + \mathbf{u} \times \mathbf{B})_x, \quad (9.190)$$

i.e. the x -component of Eq. (144). Since other Cartesian coordinates participate in Eq. (184) in a similar way, it is evident that the Lagrangian equations of motion along other coordinates yield other components of the same vector equation of motion.

So, Eq. (183) does indeed give the correct Lagrangian function, and we can use it for the further analysis, in particular to discuss the first of Eqs. (186). This relation shows that in the electromagnetic field, the generalized momentum corresponding to particle's coordinate x is *not* $p_x = m\gamma u_x$, but⁶⁰

$$P_x \equiv \frac{\partial \mathcal{L}}{\partial u_x} = p_x + qA_x. \quad (9.191)$$

Thus, as was already mentioned in brief in Sec. 6.3, particle's motion in a field may be described by two momentum vectors: the *kinetic momentum* \mathbf{p} , defined by Eq. (70), and the *canonical* (or “conjugate”) *momentum*⁶¹

Particle's
canonical
momentum

$$\mathbf{P} = \mathbf{p} + q\mathbf{A}. \quad (9.192)$$

In order to facilitate the discussion of this notion, let us generalize expression (72) for the Hamiltonian function \mathcal{H} of a free particle to the case of a particle in the field:

$$\mathcal{H} = \mathbf{P} \cdot \mathbf{u} - \mathcal{L} = (\mathbf{p} + q\mathbf{A}) \cdot \mathbf{u} - \left(-\frac{mc^2}{\gamma} + q\mathbf{u} \cdot \mathbf{A} - q\phi \right) = \mathbf{p} \cdot \mathbf{u} + \frac{mc^2}{\gamma} + q\phi. \quad (9.193)$$

Merging the first two terms exactly as it was done in Eq. (72), we get an extremely simple result,

$$\mathcal{H} = \gamma mc^2 + q\phi, \quad (9.194)$$

that may leave us wondering: where is the vector-potential \mathbf{A} here - and the field effects is has to describe? The resolution of this puzzle is easy: for a practical use (e.g., for the alternative derivation of the equations of motion), \mathcal{H} has to be presented as a function of particle's generalized coordinates (in the case of unconstrained motion, these may be the Cartesian components of vector \mathbf{r} that serves as an argument for potentials \mathbf{A} and ϕ), and the generalized momenta, i.e. the Cartesian components of vector \mathbf{P} (plus, generally, time). Hence, velocity u and factor γ should be eliminated from Eq. (194). This may be done using relation (192), $\gamma m\mathbf{u} = \mathbf{P} - q\mathbf{A}$. For such elimination, it is sufficient to notice that according

⁶⁰ With regrets, I have to use the same (common) notation as was used earlier for the electric polarization – which is not discussed below.

⁶¹ In Gaussian units, Eq. (192) has the form $\mathbf{P} = \mathbf{p} + q\mathbf{A}/c$.

to Eq. (193), difference $(\mathcal{H} - q\phi)$ is equal to the right-hand part of Eq. (72), so that the generalization of Eq. (78) is⁶²

$$(\mathcal{H} - q\phi)^2 = (mc^2)^2 + c^2(\mathbf{P} - q\mathbf{A})^2. \quad (9.195) \quad \text{Particle's Hamiltonian}$$

It is straightforward to verify that the Hamilton equations of motion for three Cartesian coordinates of the particle, obtained in the regular way⁶³ from this \mathcal{H} , may be merged into the same vector equation (144). In the nonrelativistic limit, the Taylor expansion of Eq. (195) to the first term in p^2 yields the following generalization of Eq. (74):

$$\mathcal{H} - mc^2 \approx \frac{p^2}{2m} + U = \frac{1}{2m}(\mathbf{P} - q\mathbf{A})^2 + U, \quad U = q\phi. \quad (9.196)$$

This expression for \mathcal{H} , and Eq. (183) for \mathcal{L} , give a clear view of the electromagnetic field effect account in analytical mechanics. The electric part of the total Lorentz force $q(\mathbf{E} + \mathbf{u} \times \mathbf{B})$ can perform work on the particle, i.e. change its kinetic energy - see Eq. (148) and its discussion. As a result, the scalar potential ϕ , whose gradient gives a contribution into \mathbf{E} , may be directly associated with potential energy $U = q\phi$. On the contrary, the magnetic component $q\mathbf{u} \times \mathbf{B}$ of the Lorentz force is always perpendicular to particle's velocity \mathbf{u} , and cannot work on it, and as a result cannot be described by a contribution to U . However, if \mathbf{A} did not participate in functions \mathcal{L} and/or \mathcal{H} at all, analytical mechanics would be unable to describe effects of magnetic field $\mathbf{B} = \nabla \times \mathbf{A}$ on particle's motion. Relations (183) and (197) show the wonderful way in which physics (or Mother Nature herself?) solves this problem: the vector-potential gives such contributions to both \mathcal{L} and \mathcal{H} (if the latter is considered, as it should be, a function of \mathbf{P} rather than \mathbf{p}) that cannot be uniquely attributed to either kinetic or potential energy, but ensure the correct equation of motion (144) in both the Lagrange and Hamilton formalisms.

I believe I still owe the reader a clear discussion of the physical sense of the canonical momentum \mathbf{P} . For that, let us consider a particle moving near a region of localized magnetic field $\mathbf{B}(\mathbf{r}, t)$, but not entering this region. If there is no electrostatic field (no other electric charges nearby), we can select such a local gauge that $\phi(\mathbf{r}, t) = 0$ and $\mathbf{A} = \mathbf{A}(t)$, so that Eq. (144) is reduced to

$$\frac{d\mathbf{p}}{dt} = q\mathbf{E} = -q \frac{d\mathbf{A}}{dt}, \quad (9.197)$$

immediately giving

$$\frac{d\mathbf{P}}{dt} = 0. \quad (9.198)$$

Hence, even if the magnetic field is changed in time, so that the induced electric field accelerates the particle, its conjugate momentum does not change. Hence \mathbf{P} is a variable more stable to magnetic field changes than its kinetic counterpart \mathbf{p} . This conclusion may be criticized because it relies on a specific gauge, and generally $\mathbf{P} \equiv \mathbf{p} + q\mathbf{A}$ is not gauge-invariant, because vector-potential \mathbf{A} isn't.⁶⁴ However, as

⁶² This relation may be also obtained from the expression for the Lorentz-invariant norm, $p^\alpha p_\alpha = (mc)^2$, of the 4-momentum (75), $p^\alpha = \{\mathcal{E}/c, \mathbf{p}\} = \{(\mathcal{H} - q\phi)/c, \mathbf{P} - q\mathbf{A}\}$.

⁶³ See, e.g., CM Sec. 10.1.

⁶⁴ The kinetic momentum $\mathbf{p} = m\mathbf{u}$ is just the usual $m\mathbf{u}$ product modified for relativistic effects, so that this variable is evidently gauge- (though not Lorentz-) invariant.

was already discussed in Sec. 5.3, integral $\oint \mathbf{A} \cdot d\mathbf{r}$ over a closed contour does not depend on the chosen gauge and equals to the magnetic flux Φ through the area limited by the contour – see Eq. (5.65). Integrating Eq. (197) over a closed trajectory of a particle (Fig. 14), and over the time of one orbit, we get

$$\Delta \oint_C \mathbf{p} \cdot d\mathbf{r} = -q\Delta\Phi, \quad \text{so that} \quad \Delta \oint_C \mathbf{P} \cdot d\mathbf{r} = 0, \quad (9.199)$$

where $\Delta\Phi$ is the change of flux during that time. This gauge-invariant result confirms the above conclusion about the stability of the canonical momentum to magnetic field variations.

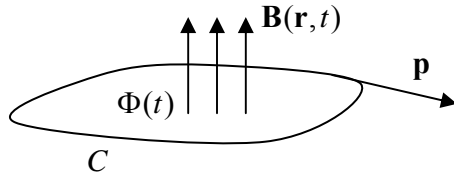


Fig. 9.14. Particle's motion around a localized magnetic flux.

Generally, Eq. (199) is invalid if a particle moves inside a magnetic field and/or changes its trajectory at the field variation. However, if the field is almost uniform, i.e. its gradient small in the sense of Eq. (177), this result is (approximately) applicable. Indeed, analytical mechanics⁶⁵ tells us that for any canonical coordinate-momentum pair $\{q_j, p_j\}$, the corresponding *action variable*,

$$J_j \equiv \frac{1}{2\pi} \oint p_j dq_j, \quad (9.200)$$

is asymptotically constant at slow variations of motion conditions. According to Eq. (191), for a particle in magnetic field, the generalized momentum corresponding to Cartesian coordinate r_j is P_j rather than p_j . Thus forming the net action variable $J \equiv J_x + J_y + J_z$, we may write

$$2\pi J = \oint \mathbf{P} \cdot d\mathbf{r} = \oint \mathbf{p} \cdot d\mathbf{r} + q\Phi = \text{const}. \quad (9.201)$$

Let us apply this relation to the motion of a nonrelativistic particle in an almost uniform magnetic field, with a small longitudinal velocity, $u_{||}/u_{\perp} \rightarrow 0$ (Fig. 15).

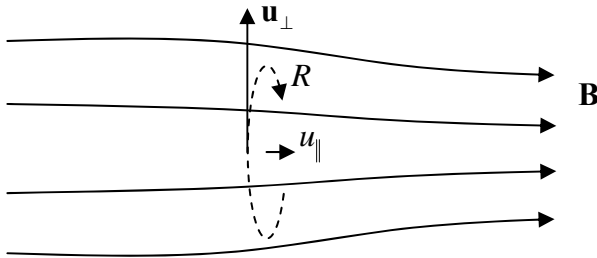


Fig. 9.15. Particle in a magnetic field with a small longitudinal gradient $\nabla B \parallel \mathbf{B}$.

In this case, Φ in Eq. (201) is the flux encircled by a cyclotron orbit, equal to $(-\pi R^2 B)$, where R is its radius given by Eq. (153), and the negative sign accounts for the fact that the “correct” direction of

⁶⁵ See, e.g., CM Sec. 10.2.

the normal vector \mathbf{n} in the definition of flux, $\Phi = \int \mathbf{B} \cdot \mathbf{n} d^2r$, is antiparallel to vector \mathbf{B} . At $u \ll c$, the kinetic momentum is just $p_{\perp} = mu_{\perp}$, while Eq. (153) yields

$$mu_{\perp} = qBR. \quad (9.202)$$

Plugging these relations into Eq. (201), we get

$$2\pi J = mu_{\perp} 2\pi R - q\pi R^2 B = m \frac{qRB}{m} 2\pi R - q\pi R^2 B = (2-1)q\pi R^2 B = -q\Phi. \quad (9.203)$$

This means that even if the circular orbit slowly moves in the magnetic field, the flux encircled by the cyclotron orbit should remain constant. One manifestation of this effect is the result already mentioned in the end of Sec. 6: if a small gradient of the magnetic field is perpendicular to the field itself, particle orbit's drift is perpendicular to ∇B , so that Φ stays constant. Now let us analyze the case of a small longitudinal gradient, $\nabla B \parallel \mathbf{B}$ (Fig. 15). If the small initial longitudinal velocity u_{\parallel} is directed toward the higher field region, in order to keep Φ constant, the cyclotron orbit has to gradually shrink. Rewriting Eq. (202) as

$$mu_{\perp} = q \frac{\pi R^2 B}{\pi R} = q \frac{|\Phi|}{\pi R}, \quad (9.204)$$

we see that this reduction of R (at constant Φ) should increase the orbiting speed u_{\perp} . But since the magnetic field cannot do work on the particle, its kinetic energy,

$$\mathcal{E} = \frac{m}{2} (u_{\parallel}^2 + u_{\perp}^2), \quad (9.205)$$

should stay constant, so that the longitudinal velocity u_{\parallel} has to decrease. Hence eventually orbit's drift has to stop, and then the orbit has to start moving back toward the region of lower fields, being essentially repulsed from the high-field region. This effect is very important, in particular, for plasma confinement: two coaxial magnetic coils, inducing magnetic fields of the same direction (Fig. 16), naturally form a “magnetic bottle” that traps charged particles injected, with sufficiently low longitudinal velocities, into the region between the coils. Such bottles are the core components of the (generally, very complex) systems used for plasma confinement, in particular in the context of the long-term efforts to achieve controllable nuclear fusion.⁶⁶

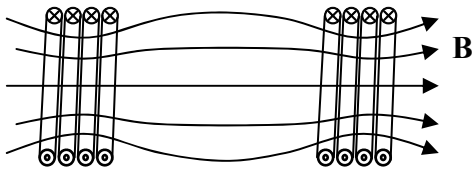


Fig. 9.16. Magnetic bottle (VERY schematically).

Returning to the constancy of magnetic flux encircled by free particles, it reminds us of the Meissner-Ochsenfeld effect discussed in Sec. 6.3, and gives a motivation for a brief revisit of the electrodynamics of superconductivity. As was emphasized in that section, superconductivity is a

⁶⁶ For the further reading on this technology, the reader may be referred, for example, to a simple monograph by F. C. Chen, *Introduction to Plasma Physics and Controllable Fusion*, vol. 1, 2nd ed., Springer, 1984, and/or a graduate-level theoretical treatment by R. D. Hazeltine and J. D. Meiss, *Plasma Confinement*, Dover, 2003.

substantially quantum phenomenon; nevertheless the notion of the conjugate momentum \mathbf{P} helps to understand its description. Indeed, the general rule of quantization of physical systems⁶⁷ is that each canonical pair $\{q_j, p_j\}$ of a generalized coordinate and the corresponding momentum is described by quantum-mechanical operators that obey the following commutation relation

$$[\hat{q}_j, \hat{p}_{j'}] = i\hbar \delta_{jj'}. \quad (9.206)$$

According to Eq. (191), for Cartesian coordinates r_j of a particle in electromagnetic field, the corresponding generalized momenta are P_j , so that their operators should obey the following commutation relations:

$$[\hat{r}_j, \hat{P}_{j'}] = i\hbar \delta_{jj'}. \quad (9.207)$$

In the coordinate representation of quantum mechanics, canonical momentum operators are described by Cartesian components of the vector operator $-i\hbar\nabla$. As a result, ignoring the rest energy mc^2 (which gives an inconsequential phase factor $\exp\{-imc^2t/\hbar\}$ in the wave function), we can use Eq. (196) to rewrite the nonrelativistic Schrödinger equation,

$$i\hbar \frac{\partial \psi}{\partial t} = \hat{\mathcal{H}}\psi, \quad (9.208)$$

as follows:

$$i\hbar \frac{\partial \psi}{\partial t} = \left(\frac{\hat{p}^2}{2m} + U \right) \psi = \left[\frac{1}{2m} (-i\hbar\nabla - q\mathbf{A})^2 + q\phi \right] \psi. \quad (9.209)$$

Thus, I believe I have finally delivered on my promise to justify the replacement (6.44) which had been used in Chapter 6 to discuss electrodynamics of superconductors, including the Meissner-Ochsenfeld effect.⁶⁸

9.8. Analytical mechanics of electromagnetic field

We have just seen that analytical mechanics of a *particle* in an electromagnetic field may be used to get some important results. The same is true for the analytical mechanics of the *field* alone, and the *field-particle system* as a whole, which will be discussed in this section. For such a space-distributed system as the field, governed by local dynamics laws (Maxwell equations), we need to apply analytical mechanics to the *local densities* ℓ and \mathcal{h} of the Lagrangian and Hamiltonian functions, defined by relations

$$\mathcal{L} = \int \ell d^3r, \quad \mathcal{H} = \int \mathcal{h} d^3r. \quad (9.210)$$

Let us start, as usual, from the Lagrange formalism. Some clue on the possible structure of the Lagrangian density ℓ may be obtained from that of the description of the particle-field interaction in this

⁶⁷ See, e.g., CM Sec. 10.1.

⁶⁸ Equation (209) is also the basis for discussion of numerous other magnetic field phenomena, including the Aharonov-Bohm and quantum Hall effects – see, e.g., QM Secs. 3.1-3.2.

formalism, which was discussed in the last section. For the case of a single particle, the interaction is described by the last two terms of Eq. (183):

$$\mathcal{L}_{\text{int}} = -q\phi - q\mathbf{u} \cdot \mathbf{A}. \quad (9.211)$$

It is obvious that if charge q is continuously distributed over some volume, we may present \mathcal{L} as a volume integral of Lagrangian density

$$\ell_{\text{int}} = -\rho\phi + \mathbf{j} \cdot \mathbf{A} = -j_{\alpha} A^{\alpha}. \quad (9.212)$$

Interaction
Lagrangian
density

Notice that the density (in contrast to \mathcal{L}_{int} itself) is Lorentz-invariant. (This is due to the contraction of the longitudinal coordinate, and hence volume, at the Lorentz transform.) Hence we may expect the density of field's Lagrangian to be Lorentz-invariant as well. Moreover, in the view of the simple, local structure of the Maxwell equations (containing only first spatial and temporal derivatives of the fields), ℓ should be a simple function of potential's 4-vector and its 4-derivative:

$$\ell = \ell(A^{\alpha}, \partial_{\alpha} A^{\beta}). \quad (9.213)$$

Also, the density should be selected in such a way that the 4-vector analog of the Lagrangian equations of motion,

$$\partial_{\alpha} \frac{\partial \ell}{\partial (\partial_{\alpha} A^{\beta})} - \frac{\partial \ell}{\partial A^{\beta}} = 0, \quad (9.214)$$

gave us correct inhomogeneous Maxwell equations (127).^{69,70} It is clear that the field part ℓ_{field} of the total Lagrangian density ℓ should be a scalar, and a quadratic form of the field strength, i.e. of $F^{\alpha\beta}$, so that the natural choice is

$$\ell_{\text{field}} = \text{const} \times F_{\alpha\beta} F^{\alpha\beta}. \quad (9.215)$$

with implied summation over both indices. Indeed, adding to this expression the interaction Lagrangian (212),

$$\ell = \ell_{\text{field}} + \ell_{\text{int}} = \text{const} \times F_{\alpha\beta} F^{\alpha\beta} - j_{\alpha} A^{\alpha}, \quad (9.216)$$

and performing differentiation, we may check that Eq. (214) indeed yields Eqs. (127), provided that the constant factor equals $(-1/4\mu_0)$.⁷¹ With that, the field Lagrangian

$$\ell_{\text{field}} = -\frac{1}{4\mu_0} F_{\alpha\beta} F^{\alpha\beta} = \frac{1}{2\mu_0} \left(\frac{E^2}{c^2} - B^2 \right) = \frac{\epsilon_0}{2} E^2 - \frac{B^2}{2\mu_0} = u_e - u_m, \quad (9.217)$$

Field's
Lagrangian
density

where u_e is the local density of the electric field energy density (1.67), and u_m is the magnetic field energy density (5.57).

⁶⁹ As a reminder, the *homogeneous* Maxwell equations (129) are satisfied by the very structure (125) of the field strength tensor.

⁷⁰ Here the implicit summation over index α plays the role similar to the convective derivative (188) in replacing the full derivative over time, in a way that reflects the symmetry of time and space in special relativity. I do not want to spend more time to justify Eq. (214) because of the reasons that will be clear very soon.

⁷¹ In the Gaussian units, the coefficient is $(-1/16\pi)$.

Let me hope the reader agrees that Eq. (217) is a wonderful result, because the Lagrangian function has the structure absolutely similar to the well-known expression $\mathcal{L} = T - U$ of the classical mechanics. So, for the field alone, the “potential” and “kinetic” energies are separable again.⁷²

As a sanity check, let us explore whether we can calculate a 4-vector analog of the Hamiltonian function \mathcal{H} . In the generic analytical mechanics,

$$\mathcal{H} = \sum_j \frac{\partial \mathcal{L}}{\partial \dot{q}_j} \dot{q}_j - \mathcal{L}. \quad (9.218)$$

However, just as for the Lagrangian function, for a field we should find the spatial density \mathcal{h} of the Hamiltonian, defined by the second of Eqs. (210), for which a natural 4-form of Eq. (218) is

$$\mathcal{h}^{\alpha\beta} = \frac{\partial \ell}{\partial (\partial_\alpha A^\gamma)} \partial^\beta A^\gamma - g^{\alpha\beta} \ell. \quad (9.219)$$

Calculated for the field alone, i.e. using Eq. (217) for ℓ , this definition yields

$$\mathcal{h}_{\text{field}}^{\alpha\beta} = \theta^{\alpha\beta} - \tau_D^{\alpha\beta}, \quad (9.220)$$

where tensor

Symmetric
energy-
momentum
tensor

$$\theta^{\alpha\beta} \equiv \frac{1}{\mu_0} \left(g^{\alpha\gamma} F_{\gamma\delta} F^{\delta\beta} + \frac{1}{4} g^{\alpha\beta} F_{\gamma\delta} F^{\gamma\delta} \right), \quad (9.221)$$

is gauge-invariant, while the remaining term,

$$\tau_D^{\alpha\beta} \equiv \frac{1}{\mu_0} g^{\alpha\gamma} F_{\gamma\delta} \partial^\delta A^\beta, \quad (9.222)$$

is not, so that it cannot correspond to any measurable variables. Fortunately, it is straightforward to verify that tensor τ_D may be presented in the form

$$\tau_D^{\alpha\beta} = \frac{1}{\mu_0} \partial_\gamma (F^{\gamma\alpha} A^\beta), \quad (9.223)$$

and as a result obeys the following relations:

$$\partial_\alpha \tau_D^{\alpha\beta} = 0, \quad \int \tau_D^{0\beta} d^3r = 0, \quad (9.224)$$

so it does not interfere with the conservation properties of the gauge-invariant, symmetric *energy-momentum tensor* (also called the *symmetric stress tensor*) $\theta^{\alpha\beta}$, to be discussed below.

Using Eqs. (125), components of the latter tensor may be expressed via the electric and magnetic fields. For $\alpha = \beta = 0$,

$$\theta^{00} = \mathcal{h}_{\text{field}} = \frac{\epsilon_0}{2} E^2 + \frac{B^2}{2\mu_0} = u_e + u_m = u, \quad (9.225)$$

⁷² Since the Lagrange equations of motion are homogeneous, the simultaneous change of sign of T and U does not change them. Thus, it is not important which of two energy densities, u_e or u_m , we count as the potential energy.

i.e. the expression for the total energy density u – see Eq. (6.104b). The other 3 components of the same row/column turn out to be just the Cartesian components of the Poynting vector, divided by c :

$$\theta^{j0} = \frac{1}{\mu_0} \left(\frac{\mathbf{E}}{c} \times \mathbf{B} \right)_j = \left(\frac{\mathbf{E}}{c} \times \mathbf{H} \right)_j = \frac{S_j}{c}, \quad \text{for } j = 1, 2, 3. \quad (9.226)$$

The remaining 9 components $\theta_{jj'}$ of the tensor, with $j' = 1, 2, 3$, are usually presented as

$$\theta^{jj'} = -\tau_{jj'}^{(M)}, \quad (9.227)$$

where $\tau^{(M)}$ is the so-called *Maxwell stress tensor*:

$$\tau_{jj'}^{(M)} = \varepsilon_0 \left(E_j E_{j'} - \frac{\delta_{jj'}}{2} E^2 \right) + \frac{1}{\mu_0} \left(B_j B_{j'} - \frac{\delta_{jj'}}{2} B^2 \right), \quad (9.228)$$

Maxwell stress tensor

so that the whole symmetric energy-momentum tensor may be conveniently presented in the following symbolic way:

$$\theta^{\alpha\beta} = \begin{pmatrix} u & \leftarrow \mathbf{S}/c & \rightarrow \\ \uparrow \mathbf{S} & & \\ \hline \frac{\mathbf{S}}{c} & -\tau_{jj'}^{(M)} & \\ \downarrow & & \end{pmatrix}. \quad (9.229)$$

The physical meaning of this tensor may be revealed in the following way. Considering Eq. (221) just as the *definition* of tensor $\theta^{\alpha\beta}$,⁷³ and using the 4-vector form of Maxwell equations, given by Eqs. (127) and (129), it is straightforward to verify an extremely simple result for the 4-derivative of the symmetric tensor:

$$\partial_\alpha \theta^{\alpha\beta} = -F^{\beta\gamma} j_\gamma. \quad (9.230)$$

Symmetric tensor's 4-derivative

This expression is valid in the presence of the electromagnetic field sources, e.g., for any system of charged particles and the field they have created. Of these 4 equations (for 4 values of index β), the temporal one (with $\beta = 0$) may be simply expressed via the energy density (225) and Poynting vector (226):

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{S} = -\mathbf{j} \cdot \mathbf{E}, \quad (9.231)$$

while 3 spatial equations (with $\beta = j = 1, 2, 3$) may be presented in the form

$$\frac{\partial}{\partial t} \frac{S_j}{c^2} - \sum_{j'=1}^3 \frac{\partial}{\partial r_{j'}} \tau_{jj'}^{(M)} = -(\rho \mathbf{E} + \mathbf{j} \times \mathbf{B})_j. \quad (9.232)$$

Integrated over a volume V limited by surface S , with the account of the divergence theorem, Eq. (231) returns us to the Poynting theorem (6.103):

⁷³ In this way, we are using Eqs. (214) and (221) just as a useful guesses, leading to the definition of $\theta^{\alpha\beta}$, and may leave their strict justification for more serious field theory courses.

$$\int_V \left(\frac{\partial u}{\partial t} + \mathbf{j} \cdot \mathbf{E} \right) d^3 r + \oint_S S_n d^2 r = 0, \quad (9.233)$$

while Eq. (232) yields:⁷⁴

$$\int_V \left[\frac{\partial}{\partial t} \frac{\mathbf{S}}{c^2} + \mathbf{f} \right] d^3 r = \sum_{j=1}^3 \oint_S \tau_{jj'}^{(M)} dA_{j'}, \quad \text{with } \mathbf{f} \equiv \rho \mathbf{E} + \mathbf{j} \times \mathbf{B}, \quad (9.234)$$

where $dA_j = n_j dA = n_j d^2 r$ is the j^{th} component of the elementary area vector $d\mathbf{A} = \mathbf{n} dA = \mathbf{n} d^2 r$ that is normal to volume's surface, and directed out of the volume – see Fig. 17.

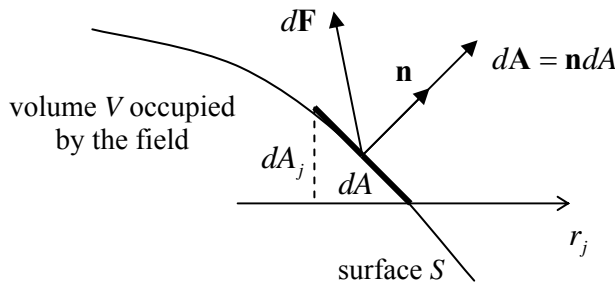


Fig. 9.17. Force $d\mathbf{F}$ exerted on a boundary element $d\mathbf{A}$ of volume V occupied by the field.

Since, according to Eq. (5.10), vector \mathbf{f} is nothing else than the density of volume-distributed forces applied from the field to the particles, we can use the 2nd Newton law, in its relativistic form (144), to rewrite Eq. (234), for a stationary volume V , as

Total
momentum's
dynamics

$$\frac{d}{dt} \left[\int_V \frac{\mathbf{S}}{c^2} d^3 r + \mathbf{p}_{\text{part}} \right] = \mathbf{F}, \quad (9.235)$$

where \mathbf{p}_{part} is the total mechanical (relativistic) momentum of all particles in volume V , and vector \mathbf{F} is defined by its Cartesian components:

Force via
the Maxwell
tensor

$$F_j = \sum_{j'=1}^3 \oint_A \tau_{jj'}^{(M)} dA_{j'}. \quad (9.236)$$

Equations (235)-(236) are our main new results. The first of them shows that vector

Electro-
magnetic
field's
momentum

$$\mathbf{g} \equiv \frac{\mathbf{S}}{c^2} \quad (9.237)$$

may be interpreted as the density of momentum of the electromagnetic field (per unit volume). This classical relation is consistent with the quantum-mechanical picture of photons being considered as ultrarelativistic particles, with momentum magnitude \mathcal{E}/c , because then the total flux of the momentum carried by photons through a unit normal area per unit time may be presented as either S_n/c or as $g_n c$. It also allows us to revisit the Poynting vector paradox that was discussed in Sec. 6.7 – see Fig. 6.9 and its

⁷⁴ Just like the Poynting theorem (233), Eq. (234) may be obtained directly from the Maxwell equations, without resorting to the 4-vector formalism – see, e.g., Sec. 8.2.2 in D. J. Griffiths, *Introduction to Electrodynamics*, 3rd ed., Prentice-Hall, 1999. However, the derivation discussed above is preferable, because it shows the wonderful unity between the laws of conservation of energy and momentum.

discussion. As has been emphasized at this discussion, vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$ in this case does not correspond to any measurable energy flow. However, the corresponding momentum (237) of the field is not only real, but may be measured by the recoil impulse⁷⁵ it gives to the field sources (say, to a magnetic coil inducing field \mathbf{H} and to the capacitor plates creating field \mathbf{E}).

Now let us turn to our second result, Eq. (236). It tells us that the 3×3 -element Maxwell stress tensor complies with the general definition of the stress tensor⁷⁶ characterizing force \mathbf{F} exerted by external forces on the boundary of a volume, in this case occupied by the electromagnetic field (Fig. 17).⁷⁷ Let us use this important result to analyze two simple examples for static fields.

(i) *Electrostatic field's effect on a perfect conductor.* Since Eq. (235) has been derived for a free space region, we have to select volume V outside the conductor, but we may align one of its faces with conductor's surface (Fig. 18).

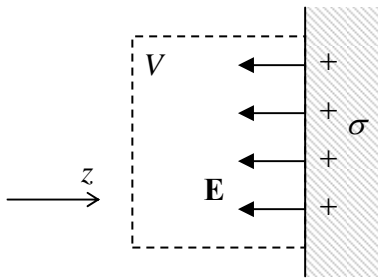


Fig. 9.18. Electrostatic field near conductor's surface.

From Chapter 2, we know that electrostatic field has to be perpendicular to conductor's surface. Selecting axis z in this direction, we have $E_x = E_y = 0$, $E_z = \pm E$, so that only diagonal components of tensor (228) do not equal zero:

$$\tau_{xx}^{(M)} = \tau_{yy}^{(M)} = -\frac{\epsilon_0}{2} E^2, \quad \tau_{zz}^{(M)} = \frac{\epsilon_0}{2} E^2, \quad (9.239)$$

Since the elementary surface area vector has just one nonvanishing component, dA_z , according to Eq. (236), only the last component (that is positive regardless of the sign of E) gives a contribution to the surface force \mathbf{F} . We see that the force exerted *by the conductor* (and eventually by external forces that hold the conductor in its equilibrium position) on the field is normal to the conductor and directed out of the field volume: $dF_z \geq 0$. Hence, by the 3rd Newton law, the force exerted *by the field* on conductor's surface is directed toward the field-filled space:

$$dF_{\text{surface}} = -dF_z = -\frac{\epsilon_0}{2} E^2 dA. \quad (9.240)$$

Electric
field's
pull

This important result could be obtained by simpler means as well. For example, one could argue, quite convincingly, that the local relation between the force and field should not depend on the global

⁷⁵ This impulse is sometimes called the *hidden momentum*; this term makes sense if the field sources have finite masses, so that their velocity change at the field variation is measurable.

⁷⁶ See, e.g., CM Sec. 7.2.

⁷⁷ Note that the field-to-particle interaction gives a vanishing contribution into the net integral, as it should for any internal interaction between internal parts of a system.

configuration creating the field, and consider a planar capacitor (Fig. 2.2) with surfaces of both plates charged by equal and opposite charges of density $\sigma = \pm \epsilon_0 E$. According to the Coulomb law, the charges should attract each other, pulling each plate toward the field region, so that Maxwell-tensor result gives the correct direction of the force. The force's magnitude (240) can be verified either by the direct integration of the Coulomb law, or by the following simple reasoning. In the plane capacitor, field $E_z = \sigma/\epsilon_0$ is equally contributed by two surface charges; hence the field created by the negative charge of the counterpart plate (not shown in Fig. 18) is $E_- = \sigma/2\epsilon_0$, and the force it exerts of the elementary surface charge $dQ = \sigma dA$ of the positively charged plate is $dF = E_- dQ = \sigma^2 dA/2\epsilon_0 = \epsilon_0 E^2 dA/2$, in accordance with Eq. (240).⁷⁸

Quantitatively, even for such high electric field as $E = 3 \text{ MV/m}$ (close to the electric breakdown in air), the “negative pressure” (dF/dA) given by Eq. (240) is of the order of 500 Pa (N/m^2), i.e. below one thousandth of the ambient atmospheric pressure (1 bar $\approx 10^5$ Pa). Still, these forces may be substantial in some cases, especially in good dielectrics (such as high-quality SiO_2 , grown at high temperature, which is broadly used in integrated circuits) that can withstand fields up to $\sim 10^9 \text{ V/m}$.

(ii) *Static magnetic field's effect on its source*⁷⁹ – say, solenoid's wall or superconductor's surface (Fig. 19).

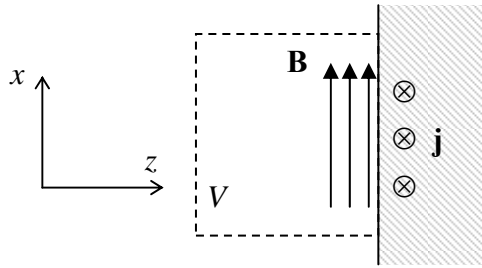


Fig. 9.19. Static magnetic field near a current-carrying surface.

With the choice of coordinates shown in Fig. 19, we have $B_x = \pm B$, $B_y = B_z = 0$, so that the Maxwell stress tensor (228) is diagonal again:

$$\tau_{xx}^{(M)} = \frac{1}{2\mu_0} B^2, \quad \tau_{yy}^{(M)} = \tau_{zz}^{(M)} = -\frac{1}{2\mu_0} B^2. \quad (9.241)$$

However, but since for this geometry only dA_z differs from 0 in Eq. (236), the sign of the resulting force is opposite to that in electrostatics: $dF_z \leq 0$, and the force exerted by the magnetic field upon the conductor's surface,

$$dF_{\text{surface}} = -dF_z = \frac{1}{2\mu_0} B^2 dA, \quad (9.242)$$

Magnetic
field's
push

⁷⁸ By the way, repeating these arguments for a plane capacitor filled with a linear dielectric, we may readily see that Eq. (240) may be generalized for this case by replacing ϵ_0 for ϵ . The similar replacement ($\mu_0 \rightarrow \mu$) is valid for Eq. (242) in a linear magnetic medium.

⁷⁹ The causal relation is not important here. Especially in the case of a superconductor, the magnetic field may be induced by another source, with the surface supercurrent \mathbf{j} just shielding the superconductor's bulk from its penetration – see Sec. 6.

corresponds to a positive pressure. For good laboratory magnets ($B \sim 10$ T), this pressure is of the order of 4×10^7 Pa ≈ 400 bars, i.e. is very substantial, so the magnets require solid mechanical design.

The direction of force (242) could be also readily predicted elementary magnetostatics arguments. Indeed, we can imagine the magnetic field volume limited by another, parallel wall with the opposite direction of surface current. According to the starting point of magnetostatics, Eq. (5.1), such surface currents of opposite directions have to repulse each other – doing that via the magnetic field.

Another explanation of the fundamental sign difference between the electric and magnetic field pressures may be provided on the electric circuit language. As we know from Chapter 2, the potential energy of the electric field stored in a capacitor may be presented in two equivalent forms,

$$U_e = \frac{CV^2}{2} = \frac{Q^2}{2C}. \quad (9.243)$$

Similarly, the magnetic field energy of in an inductive coil is

$$U_m = \frac{LI^2}{2} = \frac{\Phi^2}{2L}. \quad (9.244)$$

If we do not want to consider the work of external sources on a virtual change of the system dimensions, we should use the latter forms of these relations, i.e. consider a galvanically detached capacitor ($Q = \text{const}$) and an externally-shortcd inductance ($\Phi = \text{const}$).⁸⁰ Now if we let the electric field forces (240) drag capacitor's plates in the direction they “want”, i.e. toward each other, this would lead to a *reduction* of the capacitor thickness, and hence to an *increase* of capacitance C , and hence to a *decrease* of U_e . Similarly, for a solenoid, allowing pressure (242) to move its walls would lead to an *increase* of the solenoid volume, and hence of its inductance L , so that the potential energy U_m would be also *reduced* – as it should be. It is remarkable (actually, beautiful) how do the local field formulas (240) and (242) “know” about these global circumstances.

Finally, let us see whether the major results (237) and (242), obtained in this section, match each other. For that, let us return to the normal incidence of a plane, monochromatic wave from free space on the plane surface of a perfect conductor (see Fig. 7.8 and its discussion), and use those results to calculate the time average of pressure dF_{surface}/dA imposed by the wave on the surface. At elastic reflection from conductor's surface, electromagnetic field's momentum retains its amplitude but changes its sign, so that the momentum transferred to a unit area of the surface (i.e. average pressure) is

$$\overline{\frac{dF_{\text{surface}}}{dA}} = 2cg_{\text{incident}} = 2c \overline{\frac{S_{\text{incident}}}{c^2}} = 2c \frac{1}{c^2} \frac{E_\omega H_\omega^*}{2} = \frac{E_\omega H_\omega^*}{c}, \quad (9.245)$$

where E_ω and H_ω are complex amplitudes of the incident wave. Using relation (7.7) between these amplitudes (for $\varepsilon = \varepsilon_0$ and $\mu = \mu_0$ giving $E_\omega = cB_\omega$), we get

$$\overline{\frac{dF_{\text{surface}}}{dA}} = \frac{1}{c} cB_\omega \frac{B_\omega^*}{\mu_0} = \frac{|B_\omega|^2}{\mu_0}. \quad (9.246)$$

⁸⁰ Of course, this condition may hold “forever” only for solenoids with superconducting wiring, but even in normal-metal solenoids with practicable inductances, the flux relaxation constants L/R may be rather large (practically, up to a few minutes), quite sufficient to carry out force measurements..

On the other hand, as was discussed in Sec. 7.4, at the surface of the perfect mirror the electric field vanishes while the magnetic field doubles, so that we can use Eq. (242) with $B \rightarrow B(t) = 2\text{Re}[B_\omega e^{-i\omega t}]$. Averaging the pressure over time, we get

$$\overline{\frac{dF_{\text{surface}}}{dA}} = \frac{1}{2\mu_0} \overline{(2\text{Re}[B_\omega e^{-i\omega t}])^2} = \frac{|B_\omega|^2}{\mu_0}, \quad (9.247)$$

i.e. the same result as Eq. (246).

For the physics intuition development, it is useful to estimate the electromagnetic radiation pressure's magnitude. Even for the relatively high wave intensity S_n of 1 kW/m^2 (close to that of the direct sunlight at Earth's orbit), pressure $2cg_n = 2S_n/c$ is somewhat below $10^{-5} \text{ Pa} \sim 10^{-10} \text{ bar}$. Still, this extremely small effect was experimentally observed (by P. Lebedev) as early as in 1899, giving one of the most important confirmations of Maxwell's theory.

9.9. Exercise problems

9.1. Use the nonrelativistic Doppler effect picture to derive Eq. (4).

9.2. Show that two successive Lorentz space/time transforms in the same direction, with velocities u' and v , are equivalent to a single transform with velocity u given by Eq. (25).

9.3. $N + 1$ reference frames, numbered by index n (taking values $0, 1, \dots, N$), move in the same direction as a particle. Express the particle's velocity in frame $n = 0$ via its velocity u_N in frame number N and the set of velocities v_n of frame number n relative to the frame number $(n - 1)$.

9.4. A spaceship, moving with constant velocity v directly from the Earth, sends back brief flashes of light with period Δt_s - as measured by spaceship's clock. Calculate the period with which Earth's observers receive the signals - as measured by Earth's clock.

9.5. From the point of view of reference frame $0'$, a straight rod, parallel to axis x' , is moving, without rotation, with constant velocity \mathbf{u}' directed along axis y' . The reference frame $0'$ is itself moving relative to another ("lab") reference frame 0 , with similarly oriented axes, with a constant velocity \mathbf{v} along axis x , also without rotation - see Fig. on the right. Calculate:

- (i) the direction of rod's velocity, and
- (ii) the orientation of the rod on the $[x, y]$ plane,

as observed from the lab reference frame. Is the velocity perpendicular to the rod?

9.6. A relativistic particle moving with velocity u decays into two particles with zero rest mass.

- (i) Calculate the smallest possible angle between the decay product velocities (in the lab frame).
- (ii) What is the largest possible energy of one product particle?

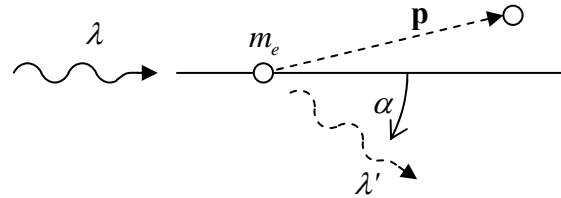
9.7. Starting from the rest at $t = 0$, a spaceship moves with a constant acceleration, as measured in its instantaneous rest frame. Find its displacement $x(t)$ from the starting point, as measured from the lab frame, and interpret the result.

9.8. Calculate the first relativistic correction to the frequency of a harmonic oscillator as a function of its amplitude.

9.9. A particle with rest mass m decays into two particles, with rest masses m_1 and m_2 . Calculate the total energy of the first decay product, in the rest frame of the decayed particle.

9.10. A relativistic particle, propagating with velocity \mathbf{v} outside of external fields, decays into two photons.⁸¹ Calculate the angular dependence of the probability of photon detection.

9.11. Photon with wavelength λ is scattered by an electron, initially at rest. Considering the photon as an ultrarelativistic particle (with the rest mass $m = 0$), find wavelength λ' of the scattered photon as a function of the scattering angle α - see Fig. on the right.⁸²



9.12. Calculate the threshold energy of a γ -photon for the reaction

$$\gamma + p \rightarrow p + \pi^0,$$

if the proton was initially at rest.

Hint: For protons $m_p c^2 \approx 938$ MeV, while for neutral pions $m_\pi c^2 \approx 135$ MeV.

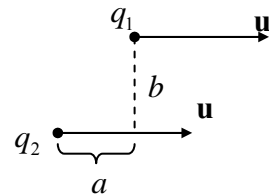
9.13. A relativistic particle with energy \mathcal{E} and rest mass m collides with a similar particle, initially at rest in the laboratory frame. Find:

- (i) the final velocity of the center of mass of the system, in the lab frame,
- (ii) the total energy of the system, in the center-of-mass frame, and
- (iii) the final velocities of both particles (in the lab frame), if they move along the same direction.

9.14. A “primed” reference frame moves with the reduced velocity $\boldsymbol{\beta} \equiv \mathbf{v}/c = \mathbf{n}_x \beta$ relative to the “lab” frame. Use Eq. (109) to spell out components T'^{00} and T'^{0j} (with $j = 1, 2, 3$) of an arbitrary contravariant 4-tensor $T'^{\mu\nu}$.

9.15. Static fields \mathbf{E} and \mathbf{B} are uniform but arbitrary (both in magnitude and in direction). What should be the velocity of an inertial reference frame to have the vectors \mathbf{E}' and \mathbf{B}' , observed from that frame, parallel? Is this solution unique?

9.16. Two charged particles, moving with the same constant velocity \mathbf{u} , are offset by distance $\mathbf{R} = \{a, b\}$ (see Fig. on the right), as measured in the lab frame. Calculate the forces between the particles - also in the lab frame.



⁸¹ Such a decay may happen, for example, with a neutral pion.

⁸² This the famous *Compton scattering* problem.

9.17. Each of two very thin, long, parallel beams of electrons of the same velocity \mathbf{u} carries electric charge of density λ per unit length (as measured in the coordinate frame moving with electrons).

(i) Calculate the distribution of the electric and magnetic fields in the system (outside the beams), as measured in the lab frame.

(ii) Calculate the interaction force between the beams (per particle) and the resulting acceleration, both in the lab frame and in the system moving with the electrons. Compare the results and give a brief discussion of the comparison.

9.18. Spell out the Lorentz transform of the scalar potential and the vector potential components, and use the result to calculate the potentials of a point charge q , moving with a constant velocity \mathbf{u} , as measured in the lab reference frame.

9.19. Calculate the scalar and vector potentials created by a time-independent electric dipole \mathbf{p} , as measured in a reference frame which moves relatively to the dipole with a constant velocity \mathbf{v} , with the shortest distance (“impact parameter”) equal to b .

9.20. Calculate the scalar and vector potentials created by a time-independent magnetic dipole \mathbf{m} , as measured in a reference frame which moves relatively to the dipole with a constant velocity $\mathbf{v} \ll c$, with the shortest distance (“impact parameter”) equal to b .

9.21. Assuming that the magnetic monopole does exist and has magnetic charge g , calculate the change $\Delta\Phi$ of magnetic flux in a superconductor ring due to the passage of single monopole through it. Evaluate $\Delta\Phi$ for the monopole charge conjectured by Dirac, $g = ng_0 \equiv n(2\pi\hbar/e)$, where n is an integer; compare the result with the magnetic flux quantum Φ_0 (6.55) and discuss their relation.

9.22.* Calculate the trajectory of a relativistic particle in a uniform electrostatic field \mathbf{E} for the case of arbitrary direction of its initial velocity $\mathbf{u}(0)$, using two different approaches – one of them different from the approach used in Sec. 6 for the case $\mathbf{u}(0) \perp \mathbf{E}$.

9.23. A charged relativistic particle with velocity u performs planar cyclotron rotation in a uniform external magnetic field B . How much would the velocity and orbit radius change at a slow change of the field to a new magnitude B' ?

9.24.* Analyze the motion of a relativistic particle in uniform, mutually perpendicular fields \mathbf{E} and \mathbf{B} , for the particular case when E is *exactly* equal to cB .

9.25.* Find the law of motion of a relativistic particle in uniform, parallel, static fields \mathbf{E} and \mathbf{B} .

9.26. Neglecting relativistic effects, calculate the smallest voltage V that has to be applied between the anode and cathode of a magnetron (see Fig. 13 and its discussion) to enable electrons to reach the anode in the absence of electron-electron interactions and collisions with the residual gas molecules. You may model the cathode and anode as two coaxial round cylinders, of radii R_1 and R_2 , respectively, assume that the magnetic field \mathbf{B} , directed along their common axis, is uniform, and neglect the initial velocity of the electrons emitted by the cathode. (After the solution, estimate the validity of the last assumption for reasonable values of parameters.)

9.27. A charged, relativistic particle has been injected into a uniform electric field that oscillates in time with frequency ω . Calculate the time dependence of the particle's velocity, as observed from a lab frame.

9.28. Analyze motion of a nonrelativistic particle in a region where the electric and magnetic fields are both constant and uniform, but not necessarily parallel or perpendicular to each other.

9.29. A static distribution of electric charge in otherwise free space has created a time-independent distribution $\mathbf{E}(\mathbf{r})$ of the electric field. Use two different approaches to express the energy density u' and the Poynting vector \mathbf{S}' , as observed in a reference frame moving with constant velocity \mathbf{v} , via the components of vector \mathbf{E} . In particular, is \mathbf{S}' equal to $(-\mathbf{v}u')$?

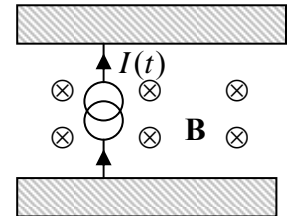
9.30. A plane wave, of frequency ω and intensity S , is normally incident on a perfect mirror, moving with velocity v in the same direction as the wave.

- (i) Calculate the reflected wave's frequency, as observed in the lab reference frame, and
 - (ii) use the Lorentz transform of the fields to calculate the reflected wave's intensity
- both as observed from the lab reference frame.

9.31. Carry out the second task of the previous problem by using the relations between wave's energy, power, and momentum.

Hint: As a byproduct, this approach should also give you the pressure exerted by the wave on the moving mirror.

9.32. Consider the simple model of plane capacitor charging by a lumped current source, shown in Fig. on the right, and prove that the momentum given by the constant, uniform external magnetic field \mathbf{B} to the current-carrying conductor is equal and opposite to the momentum of the electromagnetic field that current $I(t)$ builds up in the capacitor. (You may let the capacitor be planar and very broad, and neglect the fringe field effects.)



9.33. Consider an electromagnetic plane wave packet propagating in free space, with the electric field represented as the Fourier integral

$$\mathbf{E}(\mathbf{r}, t) = \text{Re} \int_{-\infty}^{+\infty} \mathbf{E}_k e^{i\psi_k} dk, \quad \text{with } \psi_k \equiv kz - \omega_k t, \quad \text{and } \omega_k \equiv c|k|.$$

Express the full linear momentum (per unit area of wave's front) of the packet via the complex amplitudes \mathbf{E}_k . Does the momentum depend on time? (In contrast with Problem 7.7, in this case the wave packet is not necessarily narrow.)

9.34. Calculate the pressure exerted on well-conducting walls of a waveguide with rectangular ($a \times b$) cross-section by a wave propagating along it in the fundamental (H_{10}) mode. Give an interpretation of the result.

Chapter 10. Radiation by Relativistic Charges

In this chapter, we return to the electromagnetic wave radiation by moving charges, because the review of the special relativity background in the previous chapter enables an analysis of the radiation effects for arbitrary speed of the charged particle. After an analysis of such important particular cases as synchrotron radiation and “Bremsstrahlung” (brake radiation), we will discuss the apparently unrelated effect of Coulomb losses, which nevertheless will lead us to such important phenomena as the Cherenkov radiation and transitional radiation. In the end of the chapter, I will briefly review the effects of back action of the emitted radiation on the emitting particle, whose analysis reveals some limitations of classical electrodynamics.

10.1. Liénard-Wiechert potentials

A convenient starting point for the discussion of radiation by relativistic moving charges is provided by Eqs. (8.17) for retarded potentials. In free space these formulas are reduced to

$$\phi(\mathbf{r}, t) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(\mathbf{r}', t - R/c)}{R} d^3r', \quad \mathbf{A}(\mathbf{r}, t) = \frac{\mu_0}{4\pi} \int \frac{\mathbf{j}(\mathbf{r}', t - R/c)}{R} d^3r'. \quad (10.1)$$

Here R is the magnitude of the vector,

$$\mathbf{R} = \mathbf{r} - \mathbf{r}', \quad (10.2)$$

that connects the source point \mathbf{r}' to the observation point \mathbf{r} . As a reminder, Eqs. (1) were derived from the Maxwell equations without any restrictions, and are very convenient for situations with continuous distribution of charge and current. On the other hand, for point charges, with delta-functional ρ and \mathbf{j} , it is more convenient to recast these relations into a simpler form that would not require the integration over the \mathbf{r}' space.

This reduction, however, requires care. Indeed, for a single point charge q moving with velocity \mathbf{u} , such integration of Eqs. (1), if carried out naïvely, would yield the following apparent result:

$$\phi(\mathbf{r}, t) = \frac{1}{4\pi\epsilon_0} \frac{q}{R_r}, \quad \text{i.e.} \quad \frac{\phi(\mathbf{r}, t)}{c} = \frac{\mu_0}{4\pi} \frac{qc}{R_r}; \quad \mathbf{A}(\mathbf{r}, t) = \frac{\mu_0}{4\pi} \frac{q\mathbf{u}_r}{R_r}, \quad \text{[WRONG!]} \quad (10.3)$$

where index r marks the variables to be calculated at time $t - R_r/c$. This is a good example how the science of relativity (even the special one :-)) cannot be taken too lightly. Indeed, 4-vectors (9.84)-(9.85), formed from potentials (3), would not obey the Lorentz transform rule (9.91), because distance R_r also depends on the reference frame it is measured in.

In order to correct the error, we need, first of all, to specify what exactly is R_r for a point charge. Evidently, in this case, only one space-time point $\{\mathbf{r}', t'\}$ may contribute to integrals (1) for any observation point $\{\mathbf{r}, t\}$. The point should be found from the retardation condition $t' = t - R_r/c$, i.e.

$$c(t - t') = |\mathbf{r}(t) - \mathbf{r}'(t')|. \quad (10.4)$$

Figure 1 depicts the graphical solution of this self-consistency equation as the point of intersection of the light cone of the observation point (see Fig. 9.9 and its discussion) and the trajectory of the charged

particle.¹ As in Eq. (3), I will use index r to mark all variables corresponding to the retarded point $\{\mathbf{r}', t'\}$ that satisfies Eq. (4); for example, $t' \equiv t_r$, $c(t - t_r) \equiv R_r$ (see Fig. 1), $\mathbf{u}\{\mathbf{r}', t_r\} \equiv \mathbf{u}_r$, etc, as measured in the “lab” reference frame - generally, any inertial frame that moves with the same velocity as the observation point at the moment t we are considering.

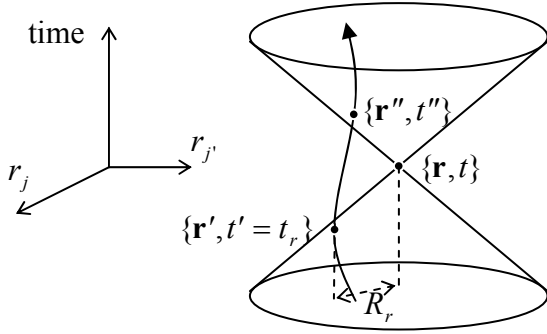


Fig. 10.1. Graphical solution of Eq. (4).

Now let us write Eqs. (1) for a point charge in another inertial reference frame $0'$, whose velocity (as measured in the lab frame) coincides, at moment t_r , with the same velocity (\mathbf{u}_r) of the point charge. In that frame the charge rests, so that

$$\phi' = \frac{q}{4\pi\epsilon_0 R'}, \quad \mathbf{A}' = 0, \quad (10.5)$$

but let us remember that this R' may not be equal to R , because the latter distance is measured in the “lab” reference frame. Let us use the identity $1/\epsilon_0 \equiv \mu_0 c^2$ to rewrite Eq. (5) in the form of components of a 4-vector similar in structure to Eq. (3):

$$\frac{\phi'}{c} = \frac{\mu_0}{4\pi} q \frac{c}{R'}, \quad \mathbf{A}' = 0. \quad (10.6)$$

Now it is easy to guess the correct answer for the whole 4-potential:

$$A^\alpha = \frac{\mu_0}{4\pi} q \frac{cu^\alpha}{u_\beta R^\beta}, \quad (10.7)$$

where (just as a reminder), $A^\alpha \equiv \{\phi/c, \mathbf{A}\}$, $u^\alpha \equiv \gamma\{c, \mathbf{u}\}$, and R^α is a 4-vector of the event distance, formed similarly to that of a single event – cf. Eq. (9.48):

$$R^\alpha \equiv \{c(t - t'), \mathbf{R}\} \equiv \{c(t - t'), \mathbf{r} - \mathbf{r}'\}. \quad (10.8)$$

Indeed, we need the 4-vector A^α that would:

- (i) obey the Lorentz transform,
- (ii) have its spatial components A_j scaling as u_j , and
- (iii) be reduced to the correct result (5) in the reference frame moving with the charge.

¹ As Fig. 1 shows, there is always another point $\{\mathbf{r}'', t''\}$, with $t'' > t$, that is formally also a solution of Eq. (4), but it does not fit Eqs. (1), because the field induced at that point would violate the causality principle.

Formula (7) evidently satisfies all these requirements, because the scalar product in its denominator is just

$$u_\beta R^\beta = \gamma \{c, -\mathbf{u}\} \cdot \{c(t-t'), \mathbf{R}\} = \gamma [c^2(t-t') - \mathbf{u} \cdot \mathbf{R}] = \gamma c(R - \boldsymbol{\beta} \cdot \mathbf{R}) = \gamma cR(1 - \boldsymbol{\beta} \cdot \mathbf{n}), \quad (10.9)$$

where $\mathbf{n} \equiv \mathbf{R}/R$ is a unit vector in the observer's direction, $\boldsymbol{\beta} \equiv \mathbf{u}/c$ is the normalized velocity of the particle, and $\gamma \equiv 1/(1 - u^2/c^2)^{1/2}$.² In the reference frame of the charge (in which $\boldsymbol{\beta} = 0$ and $\gamma = 1$), expression (9) is reduced to cR , so that Eq. (7) is correctly reduced to Eq. (6). Now let us spell out components of Eq. (7) in the lab frame (in which $t' = t_r$ and $R = R_r$):

$$\phi(\mathbf{r}, t) = \frac{1}{4\pi\epsilon_0} \frac{q}{(R - \boldsymbol{\beta} \cdot \mathbf{R})_r} = \frac{1}{4\pi\epsilon_0} q \left[\frac{1}{R(1 - \boldsymbol{\beta} \cdot \mathbf{n})} \right]_r, \quad (10.10a)$$

$$\mathbf{A}(\mathbf{r}, t) = \frac{\mu_0}{4\pi} q \left(\frac{\mathbf{u}}{R - \boldsymbol{\beta} \cdot \mathbf{R}} \right)_r = \frac{\mu_0}{4\pi} q c \left[\frac{\boldsymbol{\beta}}{R(1 - \boldsymbol{\beta} \cdot \mathbf{n})} \right]_r = \phi(\mathbf{r}, t) \frac{\mathbf{u}_r}{c^2}. \quad (10.10b)$$

These formulas are called the *Liénard-Wiechert potentials*.³ In the nonrelativistic limit, they coincide with the naïve guess (3), but in the general case include the additional factor $(1 - \boldsymbol{\beta} \cdot \mathbf{n})$ in the denominator, which describes the apparent increase of the effective charge density of the source due to the apparent change of distance R , at $\beta \sim 1$. In order to understand its origin, let us consider a simple 1D model of the radiation: a uniformly charged rod, of length l , moving directly toward an observer located at point \mathbf{r} , with a constant speed u (Fig. 2). As a result of this motion, the observer may measure the field (1) induced by the rod, within a certain time interval $[t_{\text{start}}, t_{\text{stop}}]$.

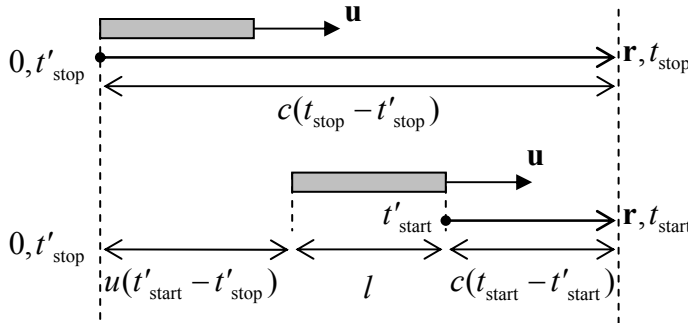


Fig. 10.2. Geometric effect behind factor $(1 - \boldsymbol{\beta} \cdot \mathbf{n})$ in the Liénard-Wiechert potentials.

That trailing end of this field pulse, observed at $t = t_{\text{stop}}$, is emitted by the far (in Fig. 2, leftmost) end of the rod at moment t'_{stop} . Due to the limited speed of the rod, $u < c$, the moment t'_{stop} comes earlier than the moment t'_{start} , at which the front end of the rod emits the field that starts the observed pulse. During the positive time interval $(t'_{\text{start}} - t'_{\text{stop}})$, the rod passes an additional distance $u(t'_{\text{start}} - t'_{\text{stop}})$ – see the bottom panel of Fig. 2. Using the evident relations shown on each of the two panels of Fig. 2 to express r , and requiring them to give the same result, we get the following relation

$$c(t_{\text{stop}} - t'_{\text{stop}}) = u(t'_{\text{start}} - t'_{\text{stop}}) + l + c(t_{\text{start}} - t'_{\text{start}}). \quad (10.11)$$

² Note the following identities: $\gamma^2 = 1/(1 - \beta^2)$ and $(\gamma^2 - 1) = \beta^2/(1 - \beta^2) = \gamma^2 \beta^2$, which may be very handy for the relativity-related algebra.

³ They were derived in 1898 by A.-M. Liénard and (apparently, independently) in 1900 by E. Wiechert.

Using it to express the difference $\Delta t'(u) \equiv t'_{\text{stop}} - t'_{\text{start}} > 0$ in the limit when $t_{\text{stop}} \rightarrow t_{\text{start}}$, i.e. when the observed radiation pulse is short, we get

$$\Delta t'(u) = \frac{l}{c-u} = \frac{l/c}{1-\beta} \equiv \frac{\Delta t(0)}{1-\beta}, \quad \text{where } \Delta t'(0) \equiv \frac{l}{c}, \quad (10.12)$$

is a factor of $1/(1-\beta)$ smaller than what it would be at negligible source speed. Hence the time interval between the retarded moments t_r for two ends of the rod is compressed as u is increased. Since the total charge of the rod does not depend on u , its linear charge density is increased, and the field in the observation point is increased accordingly. Somewhat counter-intuitively, Eq. (12) shows that this field re-normalization is independent of the source size l , and hence takes place even in the limit $l \rightarrow 0$, e.g., for a point source.⁴

So, the 4-vector formalism has provided a big help for the calculation of field potentials. Now, the electric and magnetic field corresponding to the potentials may be found by the plugging Eqs. (10) into the general formulas (6.106). This operation should be also performed very carefully, because Eqs. (6.106) require the differentiation over the coordinates $\{\mathbf{r}, t\}$ of the observation point, while we want the fields to be expressed via particle's velocity $\mathbf{u}_r \equiv (d\mathbf{r}'/dt')_r$ that participates in Eqs. (10). In order to find the relation between derivatives over t and t' , let us differentiate Eq. (4), rewritten as

$$R_r = c(t - t_r), \quad (10.13)$$

over t and t_r . In order to calculate derivative $\partial R_r / \partial t_r$, let us first differentiate identity $R^2 = \mathbf{R} \cdot \mathbf{R}$:

$$2R_r \frac{\partial R_r}{\partial t_r} = 2\mathbf{R}_r \cdot \frac{\partial \mathbf{R}_r}{\partial t_r}. \quad (10.14)$$

Since $\partial \mathbf{R}_r / \partial t_r = \partial (\mathbf{r} - \mathbf{r}') / \partial t_r = -\partial \mathbf{r}' / \partial t_r = -\mathbf{u}$, Eq. (14) yields

$$\frac{\partial R_r}{\partial t_r} = \frac{\mathbf{R}_r}{R_r} \cdot \frac{\partial \mathbf{R}_r}{\partial t_r} = -(\mathbf{n} \cdot \mathbf{u})_r. \quad (10.15)$$

Now let us differentiate the same function R_r over t , keeping \mathbf{r} fixed. On one hand, Eq. (13) yields

$$\frac{\partial R_r}{\partial t} = c - c \frac{\partial t_r}{\partial t}. \quad (10.16)$$

On the other hand, according to Eq. (4), if \mathbf{r} is fixed, t' is a function of t alone, so that, using Eq. (15), we may write

$$\frac{\partial R}{\partial t} = \frac{\partial R_r}{\partial t_r} \frac{\partial t_r}{\partial t} = -(\mathbf{n} \cdot \mathbf{u})_r \frac{\partial t_r}{\partial t}. \quad (10.17)$$

Requiring Eqs. (16) and (17) to give the same result, we get the same factor that participates in the Liénard-Wiechert potentials (10) and Eq. (12):

⁴ Note that this time compression effect (linear in β) has nothing to do with the Lorentz time dilation (9.21), which is quadratic in β . (Indeed, all our arguments above are referred to the same, lab frame.) Rather, it is close in nature to the Doppler effect.

$$\frac{\partial t_r}{\partial t} = \frac{c}{c - (\mathbf{n} \cdot \mathbf{u})_r} = \left(\frac{1}{1 - \boldsymbol{\beta} \cdot \mathbf{n}} \right)_r. \quad (10.18)$$

This relation may be readily interpreted – at least semi-quantitatively. At fixed \mathbf{r} , variation ∂t of the observation time corresponds to a small vertical shift of the light cone in Fig. 2, while ∂t_r is the corresponding shift of the retarded time t_r , i.e. of the point where the world line $\mathbf{r}'(t')$ crosses the light cone at the observation point $\mathbf{r}(t)$. It is evident from that figure that if the particle does not move (i.e. its world trajectory in a vertical straight line), then $\partial t_r = \partial t$. On the other hand, if the particles moves fast (with speed $u \approx c$) toward the observation point, its world line crosses the light cone at a small (“grazing”) angle, so that $\partial t_r \gg \partial t$, in accordance with Eq. (18).

Since the retarded time t_r , as the solution of Eq. (3), depends not only on the observation time t , but also the observation point \mathbf{r} , so we also need to calculate its spatial derivative – the gradient in \mathbf{r} -space. A calculation, absolutely similar to that carried above, yields

$$\nabla t_r = - \left[\frac{\mathbf{n}}{c(1 - \boldsymbol{\beta} \cdot \mathbf{n})} \right]_r. \quad (10.19)$$

Using Eqs. (6.106), (18) and (19), the calculation of fields from Eqs. (10) is straightforward but tedious, and is left for reader’s exercise. For the electric field, the result is:

$$\mathbf{E} = \frac{q}{4\pi\epsilon_0} \left[\frac{\mathbf{n} - \boldsymbol{\beta}}{\gamma^2(1 - \boldsymbol{\beta} \cdot \mathbf{n})^3 R^2} + \frac{\mathbf{n} \times \{(\mathbf{n} - \boldsymbol{\beta}) \times \dot{\boldsymbol{\beta}}\}}{(1 - \boldsymbol{\beta} \cdot \mathbf{n})^3 cR} \right]_r. \quad (10.20a)$$

Electric
and
magnetic
fields of
relativistic
particle

The only good news about this uncomfortably bulky result is that a similar differentiation gives essentially the same formula for the magnetic field, which may be expressed via Eq. (20a):⁵

$$\mathbf{B} = \mathbf{n}_r \times \frac{\mathbf{E}}{c}, \quad \text{i.e. } \mathbf{H} = \frac{1}{Z_0} \mathbf{n}_r \times \mathbf{E}. \quad (10.20b)$$

Thus the magnetic and electric fields are always perpendicular to each other, and related just as in a plane wave – cf. Eq. (7.6),⁶ with the only difference that now vector \mathbf{n}_r may be a function of time.

As a sanity check, let us use Eq. (20a) as an alternative way to find the electric field of a charge moving without acceleration, i.e. uniformly, along a straight line – see Fig. 9.11 (reproduced in Fig. 3) and its discussion in Sec. 5. (This example will also exhibit the challenges of practical application of the Liénard-Wiechert formulas.) In this case vector $\boldsymbol{\beta}$ does not change in time, so that the second term in Eq. (20a) vanishes, and all we need to do is to spell out the Cartesian components of the first term. Let us select the coordinate axes and time origin in the same way as shown in Fig. 3, and make a clear distinction between the actual position, $\mathbf{r}'(t) = \{ut, 0, 0\}$ of the charged particle at the instant t we are

⁵ An alternative way to derive Eqs. (20) is to plug the 4-vector of potentials, given by Eq. (7), into Eq. (9.124) to calculate the field strength tensor. This calculation yields

$$F^{\alpha\beta} = \frac{\mu_0 q}{4\pi} \frac{1}{u_\gamma R^\gamma} \frac{d}{d\tau} \left[\frac{R^\alpha u^\beta - R^\beta u^\alpha}{u_\delta R^\delta} \right].$$

Now the elements of this tensor may be identified with fields components in accordance with Eq. (9.125).

⁶ Superficially, Eq. (20b) contradicts the electrostatics where \mathbf{B} should vanish while \mathbf{E} stays finite. However, note that according to the Coulomb law for a point charge, in this case $\mathbf{E} = E\mathbf{n} = E\mathbf{n}_r$, so that $\mathbf{B} \propto \mathbf{n}_r \times \mathbf{E} \propto \mathbf{n}_r \times \mathbf{n}_r = 0$.

considering, and its retarded position $\mathbf{r}'(t_r)$, where t_r is the solution of Eq. (13), i.e. the moment when the particle's field, moving with the speed of light, reaches the observation point \mathbf{r} . In these coordinates

$$\boldsymbol{\beta} = \{\beta, 0, 0\}, \quad \mathbf{r} = \{0, 0, b\}, \quad \mathbf{r}'(t_r) = \{ut_r, 0, 0\}, \quad \mathbf{n}_r = \{\cos \theta, 0, \sin \theta\}, \quad (10.21)$$

with $\cos \theta = -ut'/R_r$, so that $[(\mathbf{n} - \boldsymbol{\beta})_x]_r = -ut'/R_r - \beta$, and for the longitudinal component of the electric field, Eq. (20a) yields

$$E_x = \frac{q}{4\pi\epsilon_0} \left[\frac{-ut_r / R - \beta}{\gamma^2 (1 - \boldsymbol{\beta} \cdot \mathbf{n})^3 R^2} \right]_r = \frac{q}{4\pi\epsilon_0} \left[\frac{-ut_r - \beta R}{\gamma^2 (1 - \boldsymbol{\beta} \cdot \mathbf{n})^3 R^3} \right]_r. \quad (10.22)$$

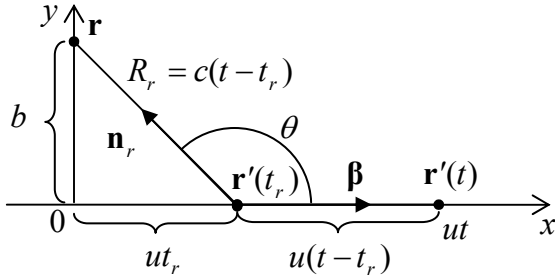


Fig. 10.3. Geometry of the linearly moving charge problem.

But according to Eq. (13), product βR_r may be presented as $\beta c(t - t_r) = u(t - t_r)$. Plugging this expression into Eq. (22), we may eliminate the explicit dependence of E_x on time t' :

$$E_x = \frac{q}{4\pi\epsilon_0} \frac{-ut}{\gamma^2 [(1 - \boldsymbol{\beta} \cdot \mathbf{n})R]_r^3}. \quad (10.23)$$

The nonvanishing transverse component of the field also has a similar form:

$$E_y = \frac{q}{4\pi\epsilon_0} \left[\frac{\sin \theta}{\gamma^2 (1 - \boldsymbol{\beta} \cdot \mathbf{n})^3 R^2} \right]_r = \frac{q}{4\pi\epsilon_0} \frac{b}{\gamma^2 [(1 - \boldsymbol{\beta} \cdot \mathbf{n})R]_r^3}, \quad (10.24)$$

while $E_z = 0$. Hence, the only combination of t_r and R_r we still need to calculate is $[(1 - \boldsymbol{\beta} \cdot \mathbf{n})R]_r$. From Fig. 3, $\boldsymbol{\beta} \cdot \mathbf{n}_r = \beta \cos \theta = -\beta ut'/R_r$, so that $(1 - \boldsymbol{\beta} \cdot \mathbf{n})R_r = R_r + \beta ut_r = c(t - t_r) + c\beta^2 t_r = ct - ct_r/\gamma^2$. What remains is to find time t_r from the self-consistency equation (13) that in our case (Fig. 3) takes the form

$$R_r^2 \equiv c^2(t - t_r)^2 = b^2 + (ut_r)^2. \quad (10.25)$$

After solving this quadratic equation (with the appropriate negative sign before the square root, in order to get $t_r < t$),

$$t_r = \gamma^2 t - \left[(\gamma^2 t)^2 - \gamma^2 (t^2 - b^2/c^2) \right]^{1/2} = \gamma^2 t - \frac{\gamma}{c} (u^2 \gamma^2 t^2 + b^2)^{1/2}, \quad (10.26)$$

we obtain a simple result:

$$[(1 - \boldsymbol{\beta} \cdot \mathbf{n})R]_r = \frac{c}{\gamma^2} (u^2 \gamma^2 t^2 + b^2)^{1/2}, \quad (10.27)$$

so that the electric field components are

$$E_x = -\frac{q}{4\pi\epsilon_0} \frac{\gamma u t}{(b^2 + \gamma^2 u^2 t^2)^{3/2}}, \quad E_y = \frac{q}{4\pi\epsilon_0} \frac{\gamma b}{(b^2 + \gamma^2 u^2 t^2)^{3/2}}, \quad E_z = 0. \quad (10.28)$$

These are exactly Eqs. (9.139),⁷ which had been obtained in Sec. 9.5 by simpler means, without the necessity to solve the self-consistency equation for t_r . However, that alternative approach was essentially based on the inertial motion of the particle, and cannot be used in problems in which particle moves with acceleration. In those problems, the second term in Eq. (20a), describing wave radiation, is essential and most important.

10.2. Radiation power

Let us calculate the angular distribution of particle's radiation. For that, we need to return to use Eqs. (20) to find the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$, and in particular its component $S_n = \mathbf{S} \cdot \mathbf{n}_r$, at large distances R from the particle. Following tradition, let us express the result as the radiated energy per unit solid angle per unit time interval dt_r of the *radiation* (rather than its *measurement*), using Eq. (18):

$$\frac{d\mathcal{P}}{d\Omega} \equiv -\frac{d\mathcal{E}}{d\Omega dt_r} = \left[R^2 S_n \frac{\partial t}{\partial t_r} \right]_r = \left[R^2 (\mathbf{E} \times \mathbf{H}) \cdot \mathbf{n} (1 - \boldsymbol{\beta} \cdot \mathbf{n}) \right]_r. \quad (10.29)$$

At sufficiently large distances from the particle, i.e. in the limit $R \rightarrow \infty$, the contribution of the first (essentially, the Coulomb field) term in the square brackets of Eq. (20a) vanishes as $1/R^2$, so that we get a key formula valid for an arbitrary law of particle motion:⁸

Angular
density of
radiation
power

$$\frac{d\mathcal{P}}{d\Omega} = \frac{Z_0 q^2}{(4\pi)^2} \frac{|\mathbf{n} \times [(\mathbf{n} - \boldsymbol{\beta}) \times \dot{\boldsymbol{\beta}}]|^2}{(1 - \mathbf{n} \cdot \boldsymbol{\beta})^5}. \quad (10.30)$$

Now, let us apply this important result to some simple cases. First of all, Eq. (30) says that a charge moving with constant velocity $\boldsymbol{\beta}$ does not radiate at all. This might be expected from our analysis of this case in Sec. 9.5, because in the reference frame moving with the charge it produces only the Coulomb electrostatic field, i.e. no radiation.

Next, let us consider a linear motion of a point charge with a nonvanishing acceleration – evidently directed along the motion line. With the coordinate axes directed as shown in Fig. 4a, each of the vectors involved in Eq. (30) has at most two nonvanishing Cartesian components:

$$\mathbf{n} = \{\sin \theta, 0, \cos \theta\}, \quad \boldsymbol{\beta} = \{0, 0, \beta\}, \quad \dot{\boldsymbol{\beta}} = \{0, 0, \dot{\beta}\}. \quad (10.31)$$

where θ is the angle between the directions of particle's motion and radiation propagation. Plugging these expressions into Eq. (30) and performing the vector multiplications, we get

$$\frac{d\mathcal{P}}{d\Omega} = \frac{Z_0 q^2}{(4\pi)^2} \dot{\beta}^2 \frac{\sin^2 \theta}{(1 - \beta \cos \theta)^5}. \quad (10.32)$$

⁷ A similar calculation of magnetic field components from Eq. (20b) gives the results identical to Eqs. (9.140).

⁸ If the direction of radiation, \mathbf{n} , does not change in time, this formula does not contain the observation point \mathbf{r} . Hence, from this point on, index r may be safely dropped for brevity, though we should always remember that $\boldsymbol{\beta}$ in Eq. (30) is the reduced velocity of the particle at the instant of radiation's *emission*, not *detection*.

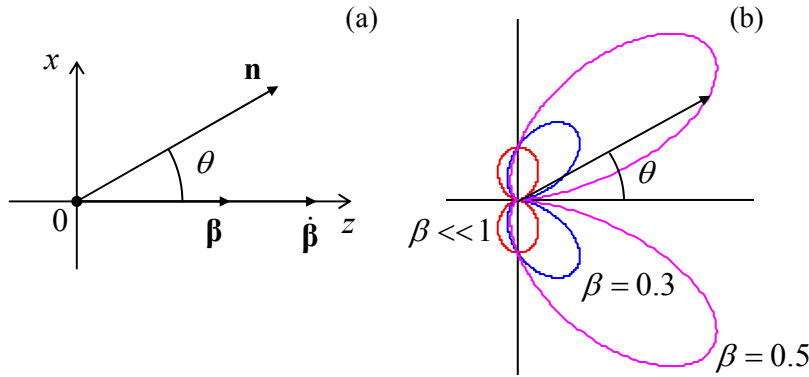


Fig. 10.4. Radiation at linear acceleration: (a) geometry of the problem, and (b) the last fraction in Eq. (32) as a function of angle θ .

Figure 4b shows the angular distribution of this radiation, for three values of particle's speed. If it is relatively low ($\beta \ll 1$), the denominator in Eq. (32) is close to 1 for all observation angles θ , so that the angular distribution of the radiation power is close to $\sin^2 \theta$ - just as it follows from the general nonrelativistic formula (8.26). However, as the velocity is increased, the denominator is less than 1 for $\theta < \pi/2$, i.e. for the forward-looking directions, and is larger than 1 for back directions. As a result, the radiation toward particle's velocity is increased (somewhat counter-intuitively, regardless of the acceleration sign!), while that in the back direction is suppressed. For ultrarelativistic particles ($\beta \rightarrow 1$), this trend is enormously exacerbated, and radiation to very small forward angles dominates. In order to describe this main part of the distribution, we may expand the trigonometric functions of θ , participating in Eq. (32), into the Taylor series in small θ , and keep only their leading terms: $\sin \theta \approx \theta$, $\cos \theta \approx 1 - \theta^2/2$, so that $(1 - \beta \cos \theta) \approx (1 + \gamma^2 \theta^2)/2\gamma^2$. The resulting expression,

$$\frac{d\mathcal{P}}{d\Omega} \approx \frac{2Z_0 q^2}{\pi^2} \dot{\beta}^2 \gamma^8 \frac{(\gamma \theta)^2}{(1 + \gamma^2 \theta^2)^5}, \quad \text{for } \gamma \gg 1, \quad (10.33)$$

describes a narrow distribution of radiation, with a maximum at angle

$$\theta_0 = \frac{1}{2\gamma} \ll 1. \quad (10.34)$$

Note that due to the axial symmetry of the result, and the fact that according to Eq. (33), $d\mathcal{P}/d\Omega = 0$ in the exact direction of particle's propagation ($\theta=0$), Eq. (40) describes a narrow circular "hollow cone" of radiation. Another important aspect of this result is how fast does the maximum radiation brightness grows with the Lorentz factor γ , i.e. with particle's energy $\mathcal{E} = \gamma mc^2$.

Still, the total radiated power \mathcal{P} (into all observation angles) at linear acceleration is not too high for any practicable values of parameters. In order to show this, it is convenient to calculate \mathcal{P} for an arbitrary motion of the particle first. It is possible to do this by a straightforward integration of Eq. (30) over the full solid angle, but let me demonstrate how \mathcal{P} may be found (or rather guessed) from the general relativistic arguments. In Sec. 8.2, we have derived Eq. (8.27) for the electric dipole radiation for nonrelativistic particle motion. That result is valid, in particular, for one charged particle whose electric dipole moment's derivative over time may be expressed as $d(q\mathbf{r})/dt = (q/m)\mathbf{p}$, where \mathbf{p} is particle's mechanical momentum (*not* its electric dipole moment). As the result, the Larmor formula (8.27) in free space, i.e. with $v = c$, reduces to

$$\mathcal{P} = \frac{Z_0}{6\pi c^2} \left(\frac{q}{m} \frac{dp}{dt} \right)^2 = \frac{Z_0 q^2}{6\pi m^2 c^2} \left(\frac{d\mathbf{p}}{dt} \cdot \frac{d\mathbf{p}}{dt} \right). \quad (10.35)$$

This is evidently not a Lorentz-invariant result, but it gives a clear hint how such an invariant, that is reduced to Eq. (35) in the nonrelativistic limit, may be formed:

$$\mathcal{P} = -\frac{Z_0 q^2}{6\pi m^2 c^2} \left(\frac{dp_\alpha}{d\tau} \cdot \frac{dp^\alpha}{d\tau} \right) = \frac{Z_0 q^2}{6\pi m^2 c^2} \left[\left(\frac{d\mathbf{p}}{d\tau} \right)^2 - \frac{1}{c^2} \left(\frac{d\mathcal{E}}{d\tau} \right)^2 \right]. \quad (10.36)$$

Plugging in the relativistic expressions, $\mathbf{p} = \gamma m c \boldsymbol{\beta}$, $\mathcal{E} = \gamma m c^2$, and $d\tau = dt/\gamma$, the last formula may be recast into the *Liénard extension* of the Larmor formula:⁹

$$\mathcal{P} = \frac{Z_0 q^2}{6\pi} \gamma^6 \left[(\dot{\boldsymbol{\beta}})^2 - (\boldsymbol{\beta} \times \dot{\boldsymbol{\beta}})^2 \right] \equiv \frac{Z_0 q^2}{6\pi} \gamma^4 \left[(\dot{\boldsymbol{\beta}})^2 + \gamma^2 (\boldsymbol{\beta} \cdot \dot{\boldsymbol{\beta}})^2 \right], \quad (10.37)$$

Total
radiation
power via $\boldsymbol{\beta}$

which may be also obtained by a direct integration of Eq. (30) over the full solid angle, thus confirming our guess. However, for some applications, it is beneficial to express \mathcal{P} the via the time evolution of particle's momentum alone. For that, we may differentiate the fundamental relativistic relation (9.78), $\mathcal{E}^2 = (mc^2)^2 + (pc)^2$, over the proper time τ to get

$$2\mathcal{E} \frac{d\mathcal{E}}{d\tau} = 2c^2 p \frac{dp}{d\tau}, \quad \text{i.e.} \quad \frac{d\mathcal{E}}{d\tau} = \frac{c^2 p}{\mathcal{E}} \frac{dp}{d\tau} = u \frac{dp}{d\tau}, \quad (10.38)$$

where, at the last transition, the magnitude of the relativistic vector relation mentioned in Chapter 9, $c^2 \mathbf{p}/\mathcal{E} = \mathbf{u}$, has been used. Plugging this relation into Eq. (36), we may rewrite it as

$$\mathcal{P} = \frac{Z_0 q^2}{6\pi m^2 c^2} \left[\left(\frac{d\mathbf{p}}{d\tau} \right)^2 - \beta^2 \left(\frac{dp}{d\tau} \right)^2 \right]. \quad (10.39)$$

Total
radiation
power via \mathbf{p}

Note the difference between the squared derivatives in this expression: in the first of them we have to differentiate the momentum vector \mathbf{p} , and only then form a scalar by squaring the resulting vector derivative, while in the second case, only the magnitude of the vector is differentiated. For example, for a circular motion with constant speed (to be analyzed in detail in the next section), the second term is zero, while the first one is not.

However, if we return to the simplest case of linear acceleration (Fig. 4), then $(d\mathbf{p}/d\tau)^2 = (dp/d\tau)^2$, and Eq. (39) is reduced to

$$\mathcal{P} = \frac{Z_0 q^2}{6\pi m^2 c^2} \left(\frac{dp}{d\tau} \right)^2 (1 - \beta^2) = \frac{Z_0 q^2}{6\pi m^2 c^2} \left(\frac{dp}{d\tau} \right)^2 \frac{1}{\gamma^2} = \frac{Z_0 q^2}{6\pi m^2 c^2} \left(\frac{dp}{dt'} \right)^2, \quad (10.40)$$

(where $t' \equiv t_r$ is the time of emitting radiation as measured as in the lab frame), i.e. formally coincides with nonrelativistic Eq. (35). In order to get a better feeling of the magnitude of this radiation, we may use the fact that $dp/dt = d\mathcal{E}/dz'$. This allows us to rewrite Eq. (40) in the following form:

⁹ The second form of Eq. (10.37), frequently more convenient for applications, may be readily obtained from the first one by applying MA Eq. (7.7a) to the vector product.

$$\mathcal{P} = \frac{Z_0 q^2}{6\pi m^2 c^2} \left(\frac{d\mathcal{E}}{dz} \right)^2 = \frac{Z_0 q^2}{6\pi m^2 c^2} \frac{d\mathcal{E}}{dz'} \frac{d\mathcal{E}}{dt'} \frac{dt'}{dz'} = \frac{Z_0 q^2}{6\pi m^2 c^2 u} \frac{d\mathcal{E}}{dz'} \frac{d\mathcal{E}}{dt'}. \quad (10.41)$$

For the most important case of ultrarelativistic motion ($u \rightarrow c$), this result may be presented as

$$\frac{\mathcal{P}}{d\mathcal{E}/dt'} \approx \frac{2}{3} \frac{d(\mathcal{E}/mc^2)}{d(z'/r_c)}, \quad (10.42)$$

where r_c is the classical radius of the particle, given by Eq. (8.41). This formula shows that the radiated power, i.e. the change of particle's energy due to radiation, is much smaller than that due to the accelerating field, unless energy as large as mc^2 is gained on the classical radius of the particle. For example, for an electron, such acceleration would require the accelerating electric field of the order of $(0.5 \text{ MV})/(3 \times 10^{-15} \text{ m}) \sim 10^{14} \text{ MV/m}$, while practicable accelerating fields are below 10^3 MV/m , limited by the electric breakdown effects. Such smallness of radiative losses of energy is actually a large advantage of linear electron accelerators - such as the famous 2-mile-long SLAC¹⁰ that can accelerate electrons or positrons to energies up to 50 GeV, i.e. to $\gamma \approx 10^5$.

10.3. Synchrotron radiation

Now let me show that in circular accelerators, the radiation is much larger. Consider a charged particle being accelerated in the direction perpendicular to its velocity \mathbf{u} (for example by the magnetic component of the Lorentz force), so that its speed u , and hence the magnitude p of its momentum, do not change. In this case, the second term in Eq. (39) vanishes, and it yields

$$\mathcal{P} = \frac{Z_0 q^2}{6\pi m^2 c^2} \left(\frac{d\mathbf{p}}{d\tau} \right)^2 = \frac{Z_0 q^2}{6\pi m^2 c^2} \left(\frac{d\mathbf{p}}{dt'} \right)^2 \gamma^2. \quad (10.43)$$

Comparing this expression with Eq. (40), we see that for the same acceleration magnitude, the electromagnetic radiation is a factor of γ^2 larger. For modern accelerators, with $\gamma \sim 10^4$ - 10^5 , such a factor creates an enormous difference. For example, if a particle is on a cyclotron orbit in a constant magnetic field (as was analyzed in Sec. 9.6), both \mathbf{u} and $\mathbf{p} = \gamma m \mathbf{u}$ obey Eq. (9.150), so that

$$\left| \frac{d\mathbf{p}}{dt'} \right| = \omega_c p = \frac{u}{R} p = \beta^2 \gamma \frac{mc^2}{R}, \quad (10.44)$$

(where R is orbit's radius), so that for the power of this *synchrotron radiation*, Eq. (43) yields

$$\mathcal{P} = \frac{Z_0 q^2}{6\pi} \beta^4 \gamma^4 \frac{c^2}{R^2}. \quad (10.45) \quad \text{Synchrotron radiation power}$$

According to Eq. (9.153), at fixed magnetic field (in particle accelerators, limited to a few Tesla produced by the beam-bending magnets), the synchrotron orbit radius R scales as γ , so that according to Eq. (45), \mathcal{P} scales as γ^2 , i.e. grows fast with particle's energy $\mathcal{E} \propto \gamma$. For example, for typical parameters of the first electron cyclotrons (such as the General Electric machine in which the synchrotron radiation was first noticed in 1947), $R \sim 1 \text{ m}$, $\mathcal{E} \sim 0.3 \text{ GeV}$ ($\gamma \sim 600$), Eq. (45) gives a very modest electron energy

¹⁰ See, e.g., <https://www6.slac.stanford.edu/>.

loss per one revolution: $\mathcal{P}\Delta t' \approx 2\pi PR/c \sim 1$ keV. However, already by the mid-1970s, electron accelerators, with $R \sim 100$ m, have reached energies $\mathcal{E} \sim 10$ GeV, and the energy loss per revolution has grown to ~ 10 MeV, becoming the major energy loss mechanism.¹¹ However, what is bad for particle accelerators and storage rings is good for the so-called *synchrotron light sources* – the electron accelerators designed specially for the generation of intensive synchrotron radiation – with the spectrum extending well beyond the visible light range. Let us now analyze the angular and spectral distributions of such radiation.

To calculate the angular distribution, let us select the coordinate axes as shown in Fig. 5, with the origin at the current location of the orbiting particle, axis z along its instant velocity (i.e. vector β), and axis x toward the orbit center.

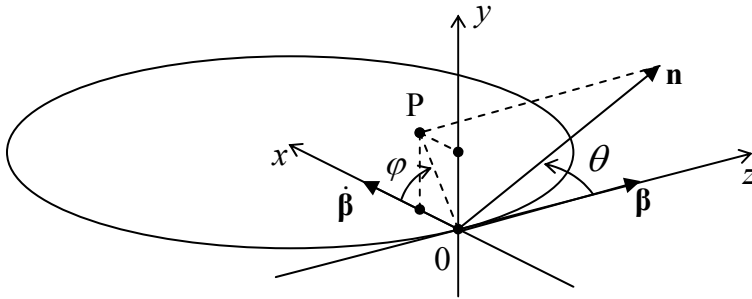


Fig. 10.5. Geometry of the synchrotron radiation problem.

In the general case, the unit vector \mathbf{n} toward the radiation observer is not within any of the coordinate planes, and hence should be described by two angles – the polar angle θ and the azimuthal angle φ between the x axis and projection OP of vector \mathbf{n} on plane $[x, y]$. Since the length of segment OP is $\sin\theta$, the Cartesian coordinates of the relevant vectors are as follows:

$$\mathbf{n} = \{\sin\theta \cos\varphi, \sin\theta \sin\varphi, \cos\theta\}, \quad \beta = \{0, 0, \beta\}, \quad \dot{\beta} = \{\dot{\beta}, 0, 0\}. \quad (10.46)$$

Plugging these coordinates into the general Eq. (30), we get

$$\frac{d\mathcal{P}}{d\Omega} = \frac{2Z_0 q^2}{\pi^2} |\dot{\beta}|^2 \gamma^6 f(\theta, \varphi), \quad \text{with } f(\theta, \varphi) \equiv \frac{1}{8\gamma^6 (1 - \beta \cos\theta)^3} \left[1 - \frac{\sin^2\theta \cos^2\varphi}{\gamma^2 (1 - \beta \cos\theta)^2} \right], \quad (10.47)$$

According to this result, just as at the linear acceleration, in the ultrarelativistic limit, most radiation goes to a narrow cone (of width $\Delta\theta \sim \gamma^{-1} \ll 1$) around vector β , i.e. around the instant direction of particle's propagation. For such small angles, and $\gamma \gg 1$, the second of Eqs. (47) is reduced to

$$f(\theta, \varphi) \approx \frac{1}{(1 + \gamma^2 \theta^2)^3} \left[1 - \frac{4\gamma^2 \theta^2 \cos^2\varphi}{(1 + \gamma^2 \theta^2)^2} \right]. \quad (10.48)$$

¹¹ For proton accelerators, such energy loss is much less of a problem, because γ of an ultrarelativistic particle (at fixed \mathcal{E}) is proportional to $1/m$, so that the estimates, at the same R , should be scaled back by $(m_p/m_e)^4 \sim 10^{13}$. Nevertheless, in the giant modern accelerators such as the LHC (with $R \approx 4.3$ km and $\mathcal{E} \approx 7$ TeV), the synchrotron radiation loss per revolution is rather noticeable ($\mathcal{P}\Delta t' \sim 6$ keV), leading not as much to particle deceleration as to substantial photoelectron emission from the beam tube walls, creating harmful defocusing effects.

Left panel of Fig. 6 shows the angular distribution $f(\theta, \varphi)$ color-coded, on the plane perpendicular to particle's instant velocity (in Fig. 5, plane $[x, y]$), while its right panel shows the intensity as a function of θ in two perpendicular directions: within the particle rotation plane (along axis x) and perpendicular to this plane (along axis y). The result shows, first of all, that, in contrast to the case of linear acceleration, the narrow radiation cone is now not hollow: the intensity maximum is reached exactly at $\theta = 0$, i.e. in particle's motion direction. Second, the radiation cone is not axially-symmetric: the intensity drops faster within the particle rotation plane (and even has nodes at $\theta = \pm 1/\gamma$).

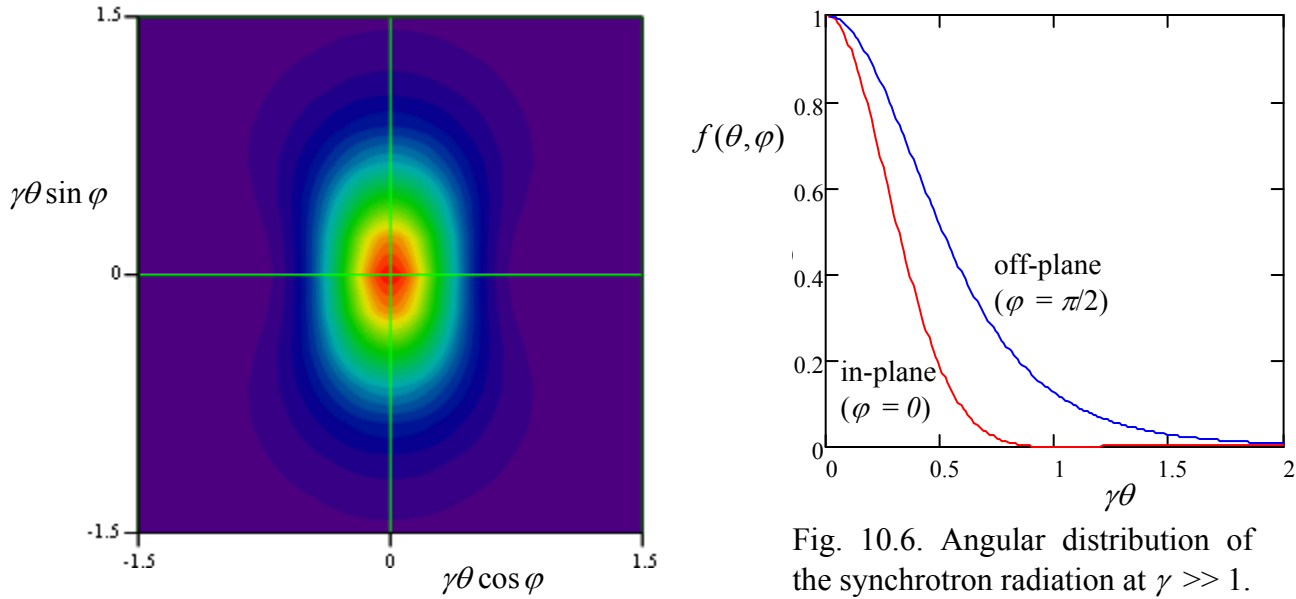


Fig. 10.6. Angular distribution of the synchrotron radiation at $\gamma \gg 1$.

Let us consider the time/frequency structure of the synchrotron radiation, now from the point of view of the observer rather than the particle itself. (In the latter picture, due to the axial symmetry of the problem, the total radiation power \mathcal{P} is evidently constant.) Its semi-quantitative picture may be obtained from the angular distribution we have just analyzed. Indeed, if an ultrarelativistic particle's radiation is observed from a point in (or close to) the rotation plane,¹² the observer is being “struck” by the narrow radiation cone once each rotation period, each “strike” giving a pulse of a short duration $\Delta t \ll \omega_c$ – see Fig. 7.

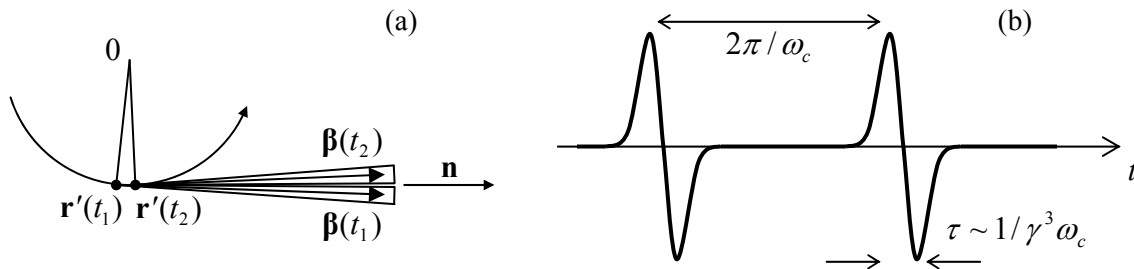


Fig. 10.7. (a) Synchrotron radiation cones at $\gamma \gg 1$, and (b) the in-plane component of their electric field, observed in the rotation plane, as a function of observation time t – schematically.

¹² If the observation point is off-plane, or if the rotation speed is much less than c , the radiation is virtually monochromatic, with frequency ω_c .

The evaluation of the time duration Δt of each pulse requires some care: its estimate $\Delta t' \sim 1/\gamma\omega_c$ is correct for the duration of the time of particle's motion while its cone is aimed at the observer. However, due to the time compression effect, discussed in detail in Sec. 1 and described by Eqs. (12) and (18), the pulse duration as seen by observer is a factor of $1/(1 - \beta)$ shorter, so that

$$\Delta t = (1 - \beta)\Delta t' \sim \frac{1 - \beta}{\gamma\omega_c} \sim \frac{1}{\gamma^3\omega_c}. \quad (10.49)$$

From the Fourier theorem, we can expect that the frequency spectrum of the radiation consists of numerous ($N \sim \gamma^3 \gg 1$) harmonics of the rotation frequency ω_c , with comparable amplitudes. However, if the orbital frequency fluctuates even slightly ($\delta\omega_c/\omega_c > 1/N \sim 1/\gamma^3$), as it happens in most practical systems, the radiation pulses are not coherent, so that the average radiation power spectrum may be calculated as that of one pulse, multiplied by number of pulses per second. In this case, the spectrum is continuous, extending from low frequencies all the way to approximately

$$\omega_{\max} \sim 1/\Delta t \sim \gamma^3\omega_c. \quad (10.50)$$

In order to verify this estimate, let us calculate the spectrum of radiation, due to a single pulse. For that, we should first make the general notion of spectrum quantitative. Let us present an arbitrary electric field (say that of the synchrotron radiation we are studying now), considered as a function of the *observation* time t (at fixed \mathbf{r}), as a Fourier integral:¹³

$$\mathbf{E}(t) = \int_{-\infty}^{+\infty} \mathbf{E}_\omega e^{-i\omega t} dt. \quad (10.51)$$

This expression may be plugged into the following formula for the total energy of the radiation pulse (i.e. of particle energy's *loss*) per unit solid angle:¹⁴

$$-\frac{d\mathcal{E}}{d\Omega} \equiv \int_{-\infty}^{+\infty} S_n(t) R^2 dt = \frac{R^2}{Z_0} \int_{-\infty}^{+\infty} |\mathbf{E}(t)|^2 dt. \quad (10.52)$$

This substitution, plus a natural change of integration order, yield

$$\frac{d\mathcal{E}}{d\Omega} = \frac{R^2}{Z_0} \int_{-\omega}^{+\omega} d\omega \int_{-\omega}^{+\omega} d\omega' \mathbf{E}_\omega \cdot \mathbf{E}_{\omega'} \int_{-\infty}^{+\infty} dt e^{-i(\omega+\omega')t}. \quad (10.53)$$

But the inner integral (over t) is just $2\pi\delta(\omega + \omega')$.¹⁵ This delta-function kills one of the frequency integrals (say, one over ω'), and Eq. (53) gives a result which may be recast as

¹³ In contrast to the single-frequency case (i.e. a monochromatic wave), we may avoid taking real part of the complex function ($\mathbf{E}_\omega e^{-i\omega t}$) if we require that $\mathbf{E}_{-\omega} = \mathbf{E}_\omega^*$. However, it is important to remember the factor $1/2$ required for the transition to a monochromatic wave of frequency ω_0 : $\mathbf{E}_\omega = \mathbf{E}_0 [\delta(\omega - \omega_0) + \delta(\omega + \omega_0)]/2$.

¹⁴ Note that the expression under the integral differs from $d\mathcal{P}/d\Omega$ defined by Eq. (29) by the absence of term $(1 - \beta \cdot \mathbf{n}) = \partial t'/\partial t$. This is natural, because this is the wave energy arriving at the observation point \mathbf{r} during time interval dt rather than dt' .

¹⁵ See, e.g. MA Eq. (14.3a).

$$-\frac{d\mathcal{E}}{d\Omega} = \int_0^{+\infty} I(\omega) d\omega, \quad \text{with } I(\omega) \equiv \frac{4\pi R^2}{Z_0} \mathbf{E}_\omega \cdot \mathbf{E}_{-\omega} = 4\pi Z_0 \varepsilon_0^2 (cR)^2 \mathbf{E}_\omega \mathbf{E}_\omega^*, \quad (10.54)$$

where the evident frequency symmetry of the scalar product $\mathbf{E}_\omega \cdot \mathbf{E}_{-\omega}$ has been utilized to fold the integral of $I(\omega)$ to positive frequencies only. The first of Eqs. (51) and the first of Eqs. (54) make the physical sense of function $I(\omega)$ clear: this is the so-called *spectral density* of the electromagnetic radiation (per unit solid angle, per unit pulse).¹⁶

In order to calculate the spectral density, we need to express function \mathbf{E}_ω via $\mathbf{E}(t)$ using the Fourier transform reciprocal to Eq. (51):

$$\mathbf{E}_\omega = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathbf{E}(t) e^{i\omega t} dt. \quad (10.55)$$

In the particular case of radiation by a single point charge, we should use the second term of Eq. (20a):

$$\mathbf{E}_\omega = \frac{1}{2\pi} \frac{q}{4\pi\varepsilon_0} \frac{1}{cR} \int_{-\infty}^{+\infty} \frac{\mathbf{n} \times \{(\mathbf{n} - \boldsymbol{\beta}) \times \dot{\boldsymbol{\beta}}\}}{(1 - \boldsymbol{\beta} \cdot \mathbf{n})^3} e^{i\omega t} dt. \quad (10.56)$$

Since vectors \mathbf{n} and $\boldsymbol{\beta}$ are natural functions of the radiation (retarded) time t' , let us use Eqs. (18) to change integration in Eq. (52) from the observation time t to time t' :

$$\mathbf{E}_\omega = \frac{q}{4\pi\varepsilon_0} \frac{1}{2\pi} \frac{1}{cR} \int_{-\infty}^{+\infty} \frac{\mathbf{n} \times \{(\mathbf{n} - \boldsymbol{\beta}) \times \dot{\boldsymbol{\beta}}\}}{(1 - \boldsymbol{\beta} \cdot \mathbf{n})^2} \exp\left\{i\omega\left(t' + \frac{R_r}{c}\right)\right\} dt'. \quad (10.57)$$

The strong inequality $R_r \gg r'$ that is implied from the beginning of this section allows us to consider the unit vector \mathbf{n} as constant and, moreover, to use approximation (8.19) to reduce Eq. (57) to

$$\mathbf{E}_\omega = \frac{1}{2\pi} \frac{q}{4\pi\varepsilon_0} \frac{1}{cR} \exp\left\{\frac{i\omega r}{c}\right\} \int_{-\infty}^{+\infty} \frac{\mathbf{n} \times \{(\mathbf{n} - \boldsymbol{\beta}) \times \dot{\boldsymbol{\beta}}\}}{(1 - \boldsymbol{\beta} \cdot \mathbf{n})^2} \exp\left\{i\omega\left(t' - \frac{\mathbf{n} \cdot \mathbf{r}'}{c}\right)\right\} dt'. \quad (10.58)$$

Plugging this expression into Eq. (54), we get¹⁷

$$I(\omega) = \frac{Z_0 q^2}{16\pi^3} \left| \int_{-\infty}^{+\infty} \frac{\mathbf{n} \times \{(\mathbf{n} - \boldsymbol{\beta}) \times \dot{\boldsymbol{\beta}}\}}{(1 - \boldsymbol{\beta} \cdot \mathbf{n})^2} \exp\left\{i\omega\left(t' - \frac{\mathbf{n} \cdot \mathbf{r}'}{c}\right)\right\} dt' \right|^2. \quad (10.59)$$

Let me remind the reader that $\boldsymbol{\beta}$ inside this integral is supposed to be taken at the retarded point $\{\mathbf{r}', t'\}$, so that Eq. (59) is fully sufficient for finding the spectral density from the law $\mathbf{r}'(t')$ of particle's motion. However, this result may be further simplified by noticing that the fraction before the exponent may be presented as a full derivative over t' ,

¹⁶ The notion of spectral density may be readily generalized to random processes – see, e.g., SM Sec. 5.4.

¹⁷ Note that for our current purposes of calculation of spectral density of radiation by a single particle, factor $\exp\{i\omega r/c\}$ has got cancelled. However, as we have seen in Chapter 8, this factor plays the central role at interference of radiation from several (many) sources. In the context of synchrotron radiation, such interference becomes important in *undulators* and *free-electron lasers* – the devices to be (qualitatively) discussed below.

$$\frac{\mathbf{n} \times \{(\mathbf{n} - \boldsymbol{\beta}) \times d\boldsymbol{\beta}/dt'\}}{(1 - \boldsymbol{\beta} \cdot \mathbf{n})^2} = \frac{d}{dt'} \left[\frac{\mathbf{n} \times (\mathbf{n} \times \boldsymbol{\beta})}{1 - \boldsymbol{\beta} \cdot \mathbf{n}} \right], \quad (10.60)$$

and working out the resulting integral by parts. At this operation, the time differentiation of the parentheses in the exponent, $d(t' - \mathbf{n} \cdot \mathbf{r}'/c)/dt' = 1 - \mathbf{n} \cdot \mathbf{u}/c = 1 - \boldsymbol{\beta} \cdot \mathbf{n}$, leads to the cancellation of denominator's remains and hence to a surprisingly simple result:¹⁸

$$I(\omega) = \frac{Z_0 q^2 \omega^2}{16\pi^3} \left| \int_{-\infty}^{+\infty} \mathbf{n} \times (\mathbf{n} \times \boldsymbol{\beta}) \exp \left\{ i\omega \left(t' - \frac{\mathbf{n} \cdot \mathbf{r}'}{c} \right) \right\} dt' \right|^2. \quad (10.61)$$

Returning to the particular case of synchrotron radiation, it is beneficial to choose the origin of time t' so that at $t' = 0$, angle θ takes its smallest value θ_0 , i.e., in terms of Fig. 5, vector \mathbf{n} is within plane $[y, z]$. Fixing this direction of axes in time, we can redraw that figure as shown in Fig. 7. In these coordinates,

$$\mathbf{n} = \{0, \sin \theta_0, \cos \theta_0\}, \quad \mathbf{r}' = \{R(1 - \cos \alpha), 0, R \sin \alpha\}, \quad \boldsymbol{\beta} \equiv \{\beta \sin \alpha, 0, \beta \cos \alpha\}, \quad (10.62)$$

where $\alpha \equiv \omega_c t'$, and an easy multiplication yields

$$\mathbf{n} \times (\mathbf{n} \times \boldsymbol{\beta}) = \beta \{\sin \alpha, \sin \theta_0 \cos \theta_0 \cos \alpha, -\sin^2 \theta_0 \sin \alpha\}, \quad (10.63)$$

$$\exp \left\{ i\omega \left(t' - \frac{\mathbf{n} \cdot \mathbf{r}'}{c} \right) \right\} = \exp \left\{ i\omega \left(t' - \frac{R}{c} \cos \theta_0 \sin \alpha \right) \right\}. \quad (10.64)$$

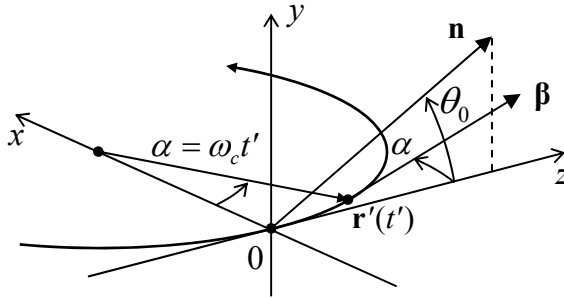


Fig. 10.7. Deriving the spectral density of synchrotron radiation. Vector \mathbf{n} is fixed in plane $[y, z]$, while vectors $\mathbf{r}'(t')$ and $\boldsymbol{\beta}(t')$ rotate in plane $[x, y]$ with angular velocity ω_c .

As we already know, in the (most interesting) ultrarelativistic limit $\gamma \gg 1$, most radiation is confined to short pulses, so that only small angles $\alpha \sim \omega_c \Delta t' \sim \gamma^{-1}$ may contribute to the integral in Eq. (61). Moreover, since most radiation goes to small angles $\theta \sim \gamma^{-1}$, it makes sense to consider only small angles $\theta_0 \sim \gamma^{-1} \ll 1$. Expanding both trigonometric functions of these small angles, participating in parentheses of Eq. (64), into Taylor series, and keeping only terms up to $O(\gamma^{-3})$, we can present them as

$$\left(t' - \frac{R}{c} \cos \theta_0 \sin \alpha \right) = \left(t' - \frac{R}{c} \omega_c t' + \frac{R}{c} \frac{\theta_0^2}{2} \omega_c t' + \frac{R}{c} \frac{\omega_c^3 t'^3}{6} \right). \quad (10.65)$$

¹⁸ Actually, this simplification is not occasional. According to Eq. (10b), the expression under the derivative is just the transverse component of the vector-potential \mathbf{A} (give or take a constant factor), and from the discussion in Sec. 8.2 we know that this component determines the electric dipole radiation of the particle (which dominates the radiation field in our current case of uncompensated electric charge).

Since $(R/c)\omega_c = u/c = \beta \approx 1$, in two last terms we may approximate this parameter by 1. However, it is crucial to distinguish the difference of two first terms, proportional to $(1 - \beta)t'$, from zero, and as we have done before we may approximate it with $t'/2\gamma^2$. In Eq. (63), which does not have such critical differences, we may be more bold, taking¹⁹

$$\mathbf{n} \times (\mathbf{n} \times \boldsymbol{\beta}) \approx \{\alpha, \theta_0, 0\} = \{\omega_c t', \theta_0, 0\}. \quad (10.66)$$

As a result, Eq. (61) is reduced to

$$I(\omega) = \frac{Z_0 q^2}{16\pi^3} |a_x \mathbf{n}_x + a_y \mathbf{n}_y|^2 = \frac{Z_0 q^2}{16\pi^3} (|a_x|^2 + |a_y|^2), \quad (10.67)$$

where a_x and a_y are the dimensionless factors,

$$\begin{aligned} a_x &\equiv \omega \int_{-\infty}^{+\infty} \omega_c t' \exp \left\{ \frac{i\omega}{2} \left((\theta_0^2 + \gamma^{-2})t' + \frac{\omega_c^2}{3} t'^3 \right) \right\} dt', \\ a_y &\equiv \omega \int_{-\infty}^{+\infty} \theta_0 \exp \left\{ \frac{i\omega}{2} \left((\theta_0^2 + \gamma^{-2})t' + \frac{\omega_c^2}{3} t'^3 \right) \right\} dt', \end{aligned} \quad (10.68)$$

Synchrotron
radiation'
spectral
density

which describe the frequency spectra of two components of the synchrotron radiation, with mutually perpendicular directions of polarization. Defining a dimensionless parameter

$$\nu \equiv \frac{\omega}{3\omega_c} (\theta_0^2 + \gamma^{-2})^{3/2}, \quad (10.69)$$

proportional to the observation frequency, and changing the integration variable to $\xi \equiv \omega_c t' / (\theta_0^2 + \gamma^{-2})^{1/2}$, integrals (68) may be reduced to the modified Bessel functions of the second kind:

$$\begin{aligned} a_x &= \frac{\omega}{\omega_c} (\theta_0^2 + \gamma^{-2}) \int_{-\infty}^{+\infty} \xi \exp \left\{ \frac{3}{2} i \nu \left(\xi + \frac{\xi^3}{3} \right) \right\} d\xi = \frac{2\sqrt{3} i}{(\theta_0^2 + \gamma^{-2})^{1/2}} \nu K_{2/3}(\nu), \\ a_y &= \frac{\omega}{\omega_c} \theta_0 (\theta_0^2 + \gamma^{-2})^{1/2} \int_{-\infty}^{+\infty} \exp \left\{ \frac{3}{2} i \nu \left(\xi + \frac{\xi^3}{3} \right) \right\} d\xi = \frac{2\sqrt{3} \theta_0}{\theta_0^2 + \gamma^{-2}} \nu K_{1/3}(\nu) \end{aligned} \quad (10.70)$$

Figure 8a shows the dependence of amplitudes a_x and a_y of the normalized observation frequency ν . It is clear that the in-plane component, proportional to a_x , is larger. (The off-plane component disappears altogether at $\theta_0 = 0$, i.e. at observation within the particle rotation plane $[x, y]$, due to the evident mirror symmetry of the problem about the plane.) It is also clear that the spectrum changes rather slowly (note the log-log scale of the plot!) until the normalized frequency, defined by Eq. (69), reaches ~ 1 . For most important observation angles $\theta_0 \sim \gamma$ this means that our estimate (50) is indeed correct, though theoretically the frequency spectrum extends to infinity.²⁰

¹⁹ By the way, this expression shows that the in-plane (x) component of the electric field is an odd function of t' (and hence t – see its sketch in Fig. 7), while the perpendicular component is an even function of time. Also notice that for an observer exactly in the rotation plane ($\theta_0 = 0$) the latter component vanishes.

²⁰ The law of the spectral density decrease at large ν may be readily obtained from the second of Eqs. (2.158) which is valid even for any (even non-integer) Bessel function index n : $a_x \propto a_y \propto \nu^{1/2} \exp\{-\nu\}$. Here the exponential factor is certainly most important.

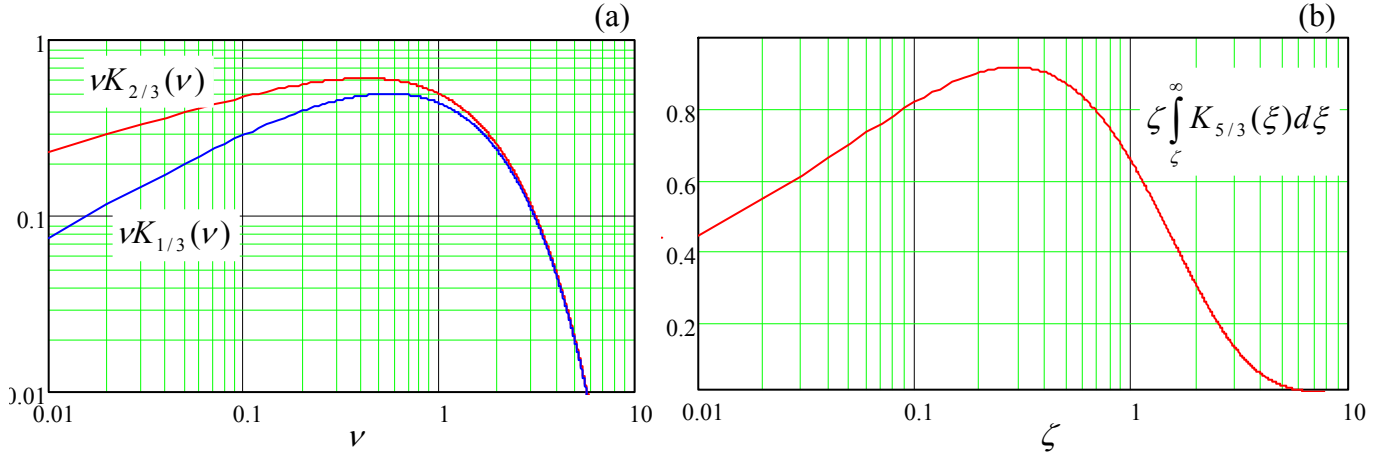


Fig. 10.8. Synchrotron radiation frequency spectra of: (a) two polarization amplitudes and (b) the total (polarization- and angle-averaged) radiation.

Naturally, a similar frequency behavior is valid for the spectral density integrated over the full solid angle. Without performing the integration,²¹ let me give the result (also valid for $\gamma \gg 1$ only) for reader's reference:

$$\oint_{4\pi} I(\omega) d\Omega = \frac{\sqrt{3}}{4\pi} q^2 \gamma \zeta \int_{\zeta}^{\infty} K_{5/3}(\xi) d\xi, \quad \zeta \equiv \frac{2}{3} \frac{\omega}{\omega_c \gamma^3}. \quad (10.71)$$

Figure 8b shows the dependence of this integral on the normalized frequency ζ . (This plot is sometimes called the “universal flux curve”.) In accordance with estimate (50), it reaches maximum at

$$\zeta_{\max} \approx 0.3, \quad \text{i.e. } \omega_{\max} \approx \frac{\omega_c}{2} \gamma^3. \quad (10.72)$$

For the new National Synchrotron Light Source (NSLS-II), that is under construction in the Brookhaven National Laboratory very close to our campus, with the ring circumference of 792 m, the electron revolution period T will be 2.64 μs . Calculating ω_c as $2\pi/T \approx 2.4 \times 10^6 \text{ s}^{-1}$, for the planned $\gamma \approx 6 \times 10^3$ ($\mathcal{E} \approx 3 \text{ GeV}$),²² we get $\omega_{\max} \sim 3 \times 10^{17} \text{ s}^{-1}$, corresponding to photon energy $\hbar\omega_{\max} \sim 200 \text{ eV}$, corresponding to soft X-rays. In the light of this estimate, the reader may be surprised by Fig. 9 that shows the projected spectra of radiation which this facility is designed to produce, with maximum photon energies up to a few keV.

The reason of this discrepancy is that in NSLS-II, and in all modern synchrotron light sources, most radiation is produced not by the circular orbit itself, but rather using special devices inserted into the electron beam path. These devices include *bend magnets* with magnetic field stronger than the average field on the orbit (which, according to Eq. (9.112), produce higher effective value of ω_c and

²¹ For that, and many other details, the interested reader may be referred, for example, to the fundamental review collection by E. E. Koch *et al.* (eds.) *Handbook on Synchrotron Radiation* (in 5 vols.), North-Holland, 1983-1991, or a more concise monograph by A. Hofmann, *The Physics of Synchrotron Radiation*, Cambridge U. Press, 2007.

²² By modern standards, this *energy* is not too high. The distinguished feature of NSLS-II is its unprecedented electron beam *intensity* (planned average beam current up to 500 mA) which should allow an extremely high synchrotron “brightness” $I(\omega)$.

hence of ω_{\max}), and *wigglers* and *undulators*: strings of several strong magnets with alternating field direction (Fig. 10), that induce periodic bending of electron trajectory, with radiation emitted at each bend.

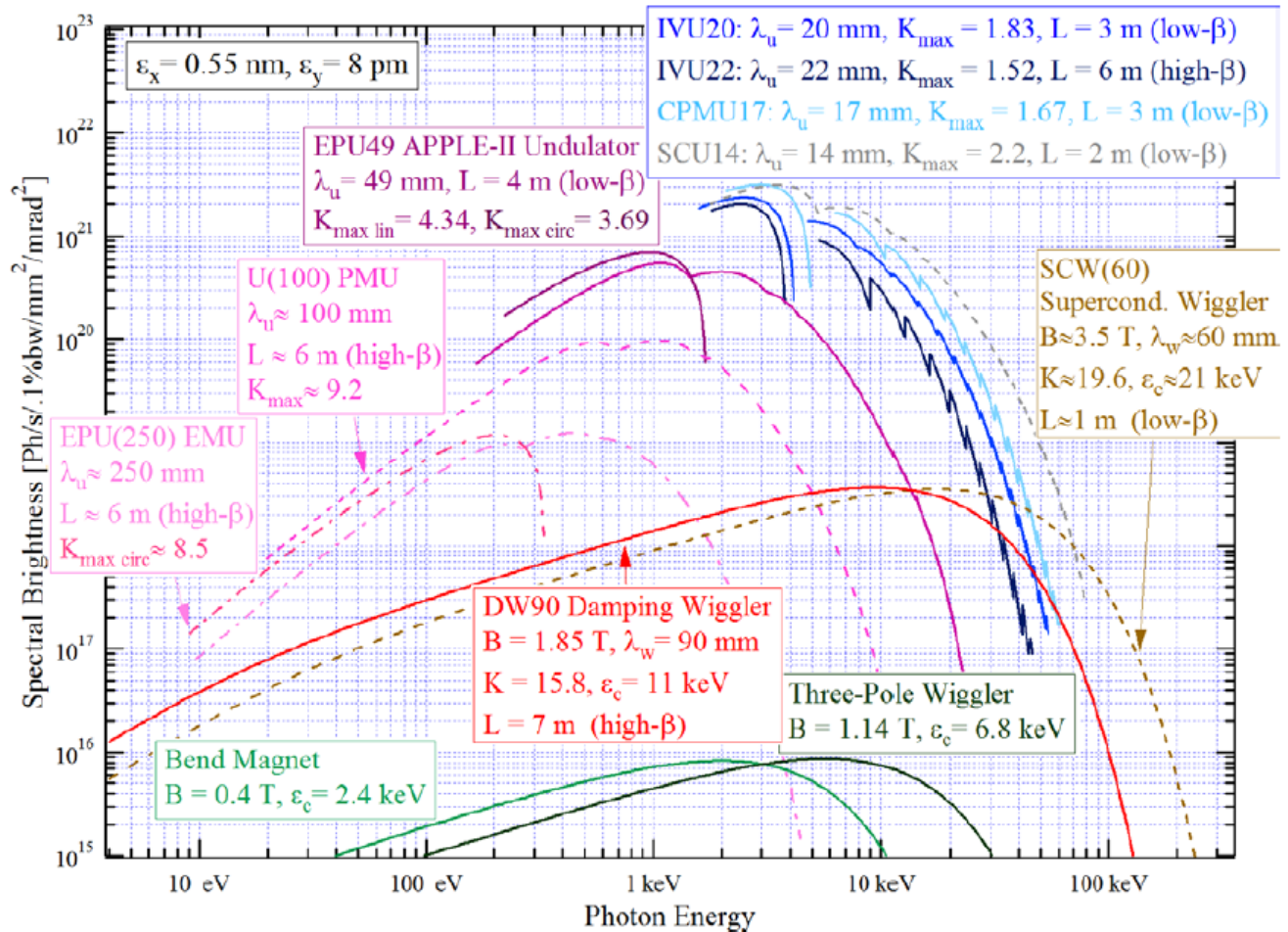


Fig. 10.9. Design brightness of various synchrotron radiation sources of the NSLS-II facility. For bend magnets and wigglers, the “brightness” may be obtained by multiplication of the spectral density $I(\omega)$ from one electron pulse, calculated above, by the number of electrons passing the source per second. (Note the non-SI units, commonly used in the synchrotron radiation community.) However, for undulators, there is an additional factor due to the partial coherence of radiation – see below. (Data from document *NSLS-II Source Properties and Floor Layout*, available online at <http://www.nsls.bnl.gov/>.)

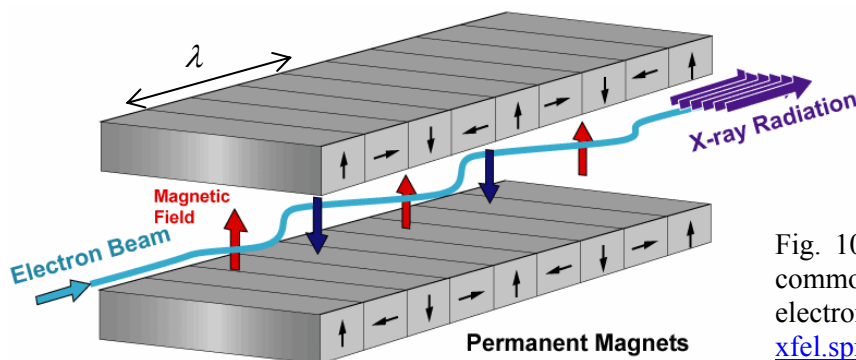


Fig. 10.10. The generic magnetic structure common for wigglers, undulators and free-electron lasers. (Adapted from <http://www.xfel.spring8.or.jp/cband/e/Undulator.htm>.)

The difference between wigglers and undulators is more quantitative than qualitative: the former devices have a larger spatial period λ (distance between the adjacent magnets of the same polarity, see Fig. 10), giving enough space for the electron beam to bend by an angle larger than γ^{-1} , i.e. larger than the radiation cone angle. As a result, the pulses radiated at each period arrive to an in-plane observer as a series of individual pulses (Fig. 11a). The shape of each pulse, and hence its frequency spectrum, are similar to those discussed above,²³ but with much higher local values of ω_c and ω_{\max} – see Fig. 9. Another difference is a much higher frequency of the peaks. Indeed, the fundamental Eq. (18) allows us to calculate the time distance between them, for the observer, as

$$\Delta t \approx \frac{\partial t}{\partial t'} \Delta t' \approx (1 - \beta) \frac{\lambda}{u} \approx \frac{1}{2\gamma^2} \frac{\lambda}{c} \ll \frac{\lambda}{c}, \quad (10.73)$$

where the first two relations are valid at $\lambda \ll R$ (the relation typically satisfied very well, see Fig. 9), and the last two relations also require the ultrarelativistic limit. As a result, the radiation intensity, that is proportional to the number of poles, is much higher than that from the bend magnets – in the NLSL-II case, more than by 2 orders of magnitude, clearly visible in Fig. 9.

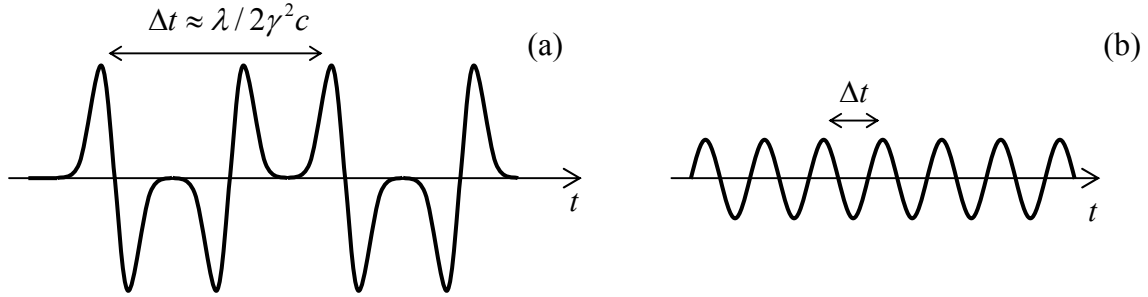


Fig. 10.11. Radiation (with in-plane polarization) from (a) a wiggler and (b) an undulator – schematically.

The situation in different in undulators – similar structures with smaller spatial period λ , in which electron's velocity vector oscillates with angular amplitude smaller than γ^{-1} . As a result, the radiation pulses overlap (Fig. 11b) and the radiation waveform is closer to sinusoidal one. As a result, the radiation spectrum narrows to the central frequency²⁴

$$\omega_0 = \frac{2\pi}{\Delta t} \approx 2\gamma^2 \frac{2\pi c}{\lambda}. \quad (10.74)$$

For example, for the LSNL-II undulators with $\lambda = 20$ mm, this formula predicts the radiation peak at photon energy $\hbar\omega_0 \approx 4$ keV, in a reasonable agreement with results of quantitative calculations, shown

²³ A small problem for the reader: use Eqs. (20) and (63) to explain the difference between shapes of pulses generated at opposite magnetic poles of the wiggler, that is schematically shown in Fig. 11a.

²⁴ This important formula may be also interpreted in the following way. Due to the relativistic length contraction (9.20), the undulator structure period as perceived by beam electrons is $\lambda' = \lambda/\gamma$, so that the central frequency of radiation is $\omega_0' = 2\pi c/\lambda' = 2\pi c\gamma/\lambda$. For the lab-frame observer, this frequency is Doppler-upshifted according to Eq. (9.44): $\omega_0 = \omega_0' [(1 + \beta)/(1 - \beta)]^{1/2} \approx 2\gamma\omega_0'$, giving the same result as Eq. (74).

in Fig. 9.²⁵ Due to the spectrum narrowing, the intensity of undulators radiation is higher than that of wigglers using the same electron beam.

This spectrum-narrowing trend is brought to its logical conclusion in the so-called *free-electron lasers*²⁶ whose basic structure is the same as that of wigglers and undulators (Fig. 10), but the radiation at each beam bend is so intense and narrow-focused that it affects the electron motion downstream the radiation cone. As a result, the radiation of all bends becomes synchronized, so its spectrum is a narrow line at frequency (70), with electromagnetic wave amplitude proportional to the number N of electrons in the structure, and hence its power proportional to N^2 (rather than to N as in wigglers and undulators).

Finally, note that wigglers, undulators, and free-electron lasers may be also used at the end of a linear electron accelerator (such as SLAC) that, as was noted above, may provide extremely high values of γ , and hence radiation frequencies (70), due to the absence of the radiation energy losses at the electron acceleration stage.

10.4. Bremsstrahlung and Coulomb losses

Surprisingly, a very similar mechanism of radiation by charged particles works at much lower spatial scale, namely at their scattering by charged particles of the propagation medium, the so-called *bremsstrahlung* - German for “brake radiation”. This effect responsible, in particular, for the continuous part of the frequency spectrum of the radiation produced by standard vacuum X-ray tubes, its incidence on a solid “anticathode”.²⁷

The bremsstrahlung in condensed matter is generally a rather complicated phenomenon, because of simultaneous involvement of many particles, and some quantum electrodynamic effect involvement. This is why I will give only a very brief glimpse at the theoretical description of this effect, for the simplest case when scattering of incoming, relatively light charged particles (such as electrons, protons, α -particles, etc.) is produced by atomic nuclei that remain virtually immobile during the scattering event (Fig. 12). This is a reasonable approximation if the energy of incoming particles is not too low, otherwise most scattering is produced by atomic electrons whose dynamics is substantially quantum – see below.

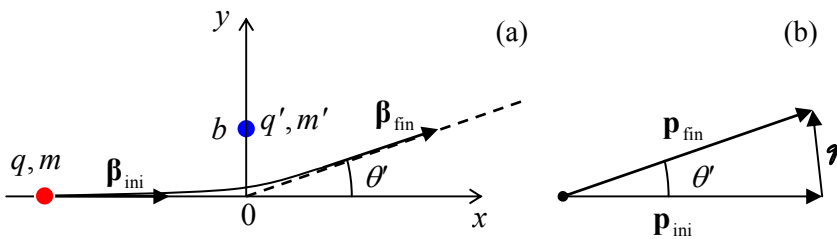


Fig. 10.12. Basic geometry of the bremsstrahlung and Coulomb loss problems in (a) direct and (b) reciprocal space.

²⁵ Much of the difference is due to the fact that those plots show the spectral density of the number of *photons* $n = \mathcal{E}/\hbar\omega$ per second, which peaks above the density of power, i.e. energy \mathcal{E} per second.

²⁶ This name is somewhat misleading, because in contrast to the usual (“quantum”) lasers, the free-electron laser operation is essentially classical and very similar to that of vacuum-tube microwave generators (such as magnetrons briefly discussed in Sec. 9.6) – see, e.g., E. Salin *et al.*, *The Physics of Free Electron Lasers*, Springer, 2000.

²⁷ Such X-ray radiation had been observed experimentally, though not correctly interpreted by N. Tesla in 1887, i.e. before the radiation was studied in detail (and much publicized) by W. Röntgen.

To calculate the frequency spectrum of radiation emitted during a single scattering event, it is convenient to use a byproduct of the last section's analysis, namely Eq. (59) with replacement (60):²⁸

$$I(\omega) = \frac{1}{4\pi^2 c} \frac{q^2}{4\pi\epsilon_0} \left| \int_{-\infty}^{+\infty} \frac{d}{dt'} \left[\frac{\mathbf{n} \times (\mathbf{n} \times \boldsymbol{\beta})}{1 - \boldsymbol{\beta} \cdot \mathbf{n}} \right] \exp \left\{ i\omega \left(t' - \frac{\mathbf{n} \cdot \mathbf{r}'}{c} \right) \right\} dt' \right|^2. \quad (10.75)$$

The typical duration τ of a single scattering event, that is described by this formula, is of the order of $a_0/c \sim (10^{-10} \text{ m})/(3 \times 10^8 \text{ m/s}) \sim 10^{-18} \text{ s}$ in solids, and only an order of magnitude longer in gases at ambient conditions. This is why for most frequencies of interest, from zero all the way up to *at least* soft X-rays,²⁹ we can use the so-called *low-frequency approximation*, taking the exponent in Eq. (75) for 1 through the whole collision event, i.e. the integration interval. This approximation immediately yields

$$I(\omega) = \frac{1}{4\pi^2 c} \frac{q^2}{4\pi\epsilon_0} \left| \frac{\mathbf{n} \times (\mathbf{n} \times \boldsymbol{\beta}_{fin})}{1 - \boldsymbol{\beta}_{fin} \cdot \mathbf{n}} - \frac{\mathbf{n} \times (\mathbf{n} \times \boldsymbol{\beta}_{ini})}{1 - \boldsymbol{\beta}_{ini} \cdot \mathbf{n}} \right|^2. \quad (10.76)$$

In the nonrelativistic limit ($\beta_{ini}, \beta_{fin} \ll 1$), this formula is reduced to³⁰

$$I(\omega) = \frac{1}{4\pi c} \frac{q^2}{4\pi^2 \epsilon_0} \frac{\boldsymbol{\varphi}^2}{m^2 c^3} \sin^2 \theta. \quad (10.77)$$

where $\boldsymbol{\varphi}$ is the momentum transferred from the scattering center to the scattered charge (Fig. 12):³¹

$$\boldsymbol{\varphi} \equiv \mathbf{p}_{fin} - \mathbf{p}_{ini} = m\Delta\mathbf{u} = mc\Delta\boldsymbol{\beta} = mc(\boldsymbol{\beta}_{fin} - \boldsymbol{\beta}_{ini}), \quad (10.78)$$

and θ is the angle between vector $\boldsymbol{\varphi}$ and the direction \mathbf{n} toward the observer.

The most important feature of result (77) is the frequency-independent (“white”) spectrum of the radiation, very typical for any rapid leaps, which may be approximated as theta-functions of time. (Note, however, that this is only valid for a fixed value of $\boldsymbol{\varphi}$, so that the statistics of this parameter, to be discussed in a minute, “colors” the radiation.) Note also the angular distribution of the radiation, forming the usual “doughnut” shape about the momentum transfer vector $\boldsymbol{\varphi}$. In particular, this means that in typical cases when $\boldsymbol{\varphi} \sim p$, the bremsstrahlung produces a significant radiation flow in the direction back to the particle source – the fact significant for the operation of X-ray tubes.

Now integrating over all wave propagation angles, just as we did for the instant radiation power in Sec. 8.2, we get the spectral density of the full energy loss,

²⁸ In publications on this topic (whose development peak was in the 1920s and 1930s), Gaussian units are more common, and letter Z is usually reserved for expressing charges as multiples the fundamental charge e , rather than for the wave impedance. This is why, in order to avoid confusion, in this section I will use $1/\epsilon_0 c \equiv Z_0$ for the free-space wave impedance and, still sticking to the same SI units as used through my lecture notes, will write the coefficients in a form that makes the transfer to the Gaussian units trivial: it is sufficient to replace all $(qq'/4\pi\epsilon_0)_{SI}$ with $(qq')_{\text{Gaussian}}$. In the (rare) cases when I spell out the charge values, I will use a different font: $q \equiv \mathcal{Z}e$, $q' \equiv \mathcal{Z}'e$.

²⁹ A more careful analysis shows that this approximation is actually quite reasonable up to much higher frequencies of the order of γ^2/τ .

³⁰ Evidently, this result (but not the general Eq. (76)!) may be derived from Eq. (8.27) as well.

³¹ Please note the font-marked difference between this variable ($\boldsymbol{\varphi}$) and particle's electric charge (q).

$$-\frac{d\mathcal{E}}{d\omega} = \oint_{4\pi} I(\omega) d\Omega = \frac{2}{3\pi c} \frac{q^2}{4\pi^2 \epsilon_0} \frac{\varphi^2}{m^2 c^3}. \quad (10.79)$$

The main new feature of bremsstrahlung (as of most scattering problems³²), is the necessity to take into account the randomness of the impact parameter b (Fig. 12). For elastic ($\beta_{ini} = \beta_{fin} \equiv \beta$) Coulomb collisions we can use the so-called Rutherford formula for the differential cross-section of scattering³³

$$\frac{d\sigma}{d\Omega'} = \left(\frac{qq'}{4\pi\epsilon_0} \right)^2 \left(\frac{1}{2pc\beta} \right)^2 \frac{1}{\sin^4(\theta'/2)}. \quad (10.80)$$

Here $d\sigma = 2\pi b db$ is the elementary area of the sample cross-section (as visible from the direction of incident particles) corresponding to particle scattering into an elementary body angle³⁴

$$d\Omega' = 2\pi \sin \theta' |d\theta'|. \quad (10.81)$$

Differentiating the geometric relation that is evident from Fig. 12,

$$\varphi = 2p \sin \frac{\theta'}{2}, \quad (10.82)$$

we may present Eq. (80) may be presented in a more convenient form

$$\frac{d\sigma}{d\varphi} = 8\pi \left(\frac{qq'}{4\pi\epsilon_0} \right)^2 \frac{1}{u^2 \varphi^3}. \quad (10.83)$$

Now combining Eqs. (79) and (83), we get

$$-\frac{d\mathcal{E}}{d\omega} \frac{d\sigma}{d\varphi} = \frac{16}{3} \frac{q^2}{4\pi\epsilon_0} \left(\frac{qq'}{4\pi\epsilon_0 mc^2} \right)^2 \frac{1}{c\beta^2} \frac{1}{\varphi}. \quad (10.84)$$

This product is called the *differential radiation cross-section*. When averaged it over all values φ (which is equivalent to averaging over all values of the impact parameter), it gives a convenient measure of radiation intensity. Indeed, after the multiplication by the volume density n of independent scattering centers, the integral gives particle's energy loss by unit bandwidth of radiation by unit path length - $d^2\mathcal{E}/d\omega dx$. A technical problem here is that the integral of $1/\varphi$ formally diverges at both infinite and zero values of φ . However, these divergences are very weak (logarithmic), and the integral converges due to virtually any reason unaccounted for by our simple analysis. The standard simple way to account for these effects is to write

$$-\frac{d^2\mathcal{E}}{d\omega dx} \approx \frac{16}{3} n \frac{q^2}{4\pi\epsilon_0} \left(\frac{qq'}{4\pi\epsilon_0 mc^2} \right)^2 \frac{1}{c\beta^2} \ln \frac{\varphi_{\max}}{\varphi_{\min}}, \quad (10.85)$$

³² See, e.g., CM Sec. 3.7.

³³ See, e.g., CM Eq. (3.72) with constant $\alpha = qq'/4\pi\epsilon_0$. In the form used in Eq. (80), the Rutherford formula is also valid for small-angle scattering of relativistic particles, the criterion being $|\Delta\beta| \ll 2/\gamma$.

³⁴ Angle θ' and differential $d\Omega'$, describing the direction of scattered *particles*, should not be confused with θ and $d\Omega$ describing directions of the *radiation* emitted at the scattering event.

and then plug, instead of ϑ_{\max} and ϑ_{\min} , scales of the most important effects limiting the small momentum range. At classical analysis, according to Eq. (82), $\vartheta_{\max} = 2p$. To estimate ϑ_{\min} , let us note that very small momentum transfer takes place when the impact parameter b is very large and hence the effective scattering time $\tau \sim b/v$ is very long. Recalling the condition of the low-frequency approximation, we may associate ϑ_{\min} with $\tau \sim 1/\omega$ and hence with $b \sim u\tau \sim v/\omega$. Since for the small scattering angles, ϑ may be estimated as the impulse $F\tau \sim (qq'/4\pi\epsilon_0 b^2)\tau$ of the Coulomb force, so that $\vartheta_{\min} \sim (qq'/4\pi\epsilon_0)\omega/u^2$, and Eq. (85) becomes

Classical
brems-
strahlung

$$-\frac{d^2\mathcal{E}}{d\omega dx} \approx \frac{16}{3} n \frac{q^2}{4\pi\epsilon_0} \left(\frac{qq'}{4\pi\epsilon_0 mc^2} \right)^2 \frac{1}{c\beta^2} \ln \left(\frac{4\pi\epsilon_0 2mu^3}{qq'\omega} \right). \quad (10.86)$$

This is *Bohr's formula* for what is called the *classical bremsstrahlung*. We see that the low momentum cutoff indeed makes the spectrum colored, with more energy going to lower frequencies. There is even a formal divergence at $\omega \rightarrow 0$; however, this divergence is integrable, so it does not present a problem in finding the total energy radiative losses ($-d\mathcal{E}/dx$) as an integral of Eq. (86) over all radiated frequencies ω . A larger problem for this procedure is the upper integration limit, $\omega \rightarrow \infty$, at which the integral diverges. This means that our approximate description, which considers the collision as an elastic process, becomes wrong, and needs to be amended by taking into account the difference between the initial and final kinetic energies of the particle due to radiation of the energy quantum $\hbar\omega$ of the emitted photon:

$$\frac{p_{ini}^2}{2m} - \frac{p_{fin}^2}{2m} = \hbar\omega. \quad (10.87)$$

As a result, taking into account that the minimum and maximum values of ϑ correspond to, respectively, the parallel and antiparallel alignments of vectors \mathbf{p}_{ini} and \mathbf{p}_{fin} , we get

Quantum
brems-
strahlung

$$\ln \frac{\vartheta_{\max}}{\vartheta_{\min}} = \ln \frac{p_{ini} + p_{fin}}{p_{ini} - p_{fin}} = \ln \frac{(p_{ini} + p_{fin})^2}{p_{ini}^2 - p_{fin}^2} = \ln \frac{[\mathcal{E}^{1/2} + (\mathcal{E} - \hbar\omega)^{1/2}]^2}{\hbar\omega}. \quad (10.88)$$

Plugged into Eq. (85), this expression yields the so-called *Bethe-Heitler formula for quantum bremsstrahlung*.³⁵ Note that at this approach, ϑ_{\max} is close to that of the classical approximation, but $\vartheta_{\min} \sim \hbar\omega/u$, so that

$$\frac{\vartheta_{\min}|_{\text{classical}}}{\vartheta_{\min}|_{\text{quantum}}} \sim \frac{\alpha z z'}{\beta}, \quad (10.89)$$

where z and z' are particles' charges in units of e , and α is the *fine structure* (“Sommerfeld”) *constant*,

$$\alpha \equiv \frac{e^2}{4\pi\epsilon_0 \hbar c} \Big|_{\text{SI}} = \frac{e^2}{\hbar c} \Big|_{\text{Gaussian}} \approx \frac{1}{137} \ll 1, \quad (10.90)$$

³⁵ The modifications of this formula necessary for the relativistic case description are surprisingly minor - see, e.g., Chapter 15 of J. D. Jackson, *Classical Electrodynamics*, 3rd ed., Wiley 1999. For more detail, the standard reference monograph on bremsstrahlung is W. Heitler, *The Quantum Theory of Radiation*, 3rd ed., Oxford U. Press 1954 (reprinted in 2010 by Dover).

which is one of the basic notions of quantum mechanics.³⁶ For most cases of practical interest, ratio (89) is smaller than 1, and since we have to keep the highest value of \mathcal{P}_{\min} , the Bethe-Heitler formula should be used.

Now nothing prevents us from calculating the total radiative losses of energy per unit length:

$$-\frac{d\mathcal{E}}{dx} = \int_0^{\infty} \left(-\frac{d^2\mathcal{E}}{d\omega dz} \right) d\omega = \frac{16}{3} n \frac{q^2}{4\pi\epsilon_0 c} \left(\frac{qq'}{4\pi\epsilon_0 mc^2} \right)^2 \frac{1}{\beta^2} 2 \int_0^{\omega_{\max}} \ln \frac{T^{1/2} - (T - \hbar\omega)^{1/2}}{(\hbar\omega)^{1/2}} d\omega, \quad (10.91)$$

where $\hbar\omega_{\max} = \mathcal{E}$ is the maximum energy of the radiation quantum. By introducing the dimensionless integration variable $\xi \equiv \hbar\omega/\mathcal{E} = 2\hbar\omega/(mu^2/2)$ this integral is reduced to the table one,³⁷ and we get

$$-\frac{d\mathcal{E}}{dx} = \frac{16}{3} n \frac{q^2}{4\pi\epsilon_0 c} \left(\frac{qq'}{4\pi\epsilon_0 mc^2} \right)^2 \frac{1}{\beta^2} \frac{u^2}{\hbar} = \frac{16}{3} n \left(\frac{q'^2}{4\pi\epsilon_0 \hbar c} \right) \left(\frac{q^2}{4\pi\epsilon_0} \right)^2 \frac{1}{mc^2}. \quad (10.92)$$

In my usual style, I would give you an estimate of the losses for a typical case; however, let me compare them to a parallel energy loss mechanism, the so-called *Coulomb losses*, due to the transfer of mechanical impulse from the scattered particle to the scattering center. (This energy eventually goes into an increase of the thermal energy of the scattering medium.) Using Eqs. (9.139) for the electric field of a linearly moving charge, we can readily find the momentum it transfers to charge q' :³⁸

$$\Delta p' = |(\Delta p')_y| = \left| \int_{-\infty}^{+\infty} (\dot{p}')_y dt \right| = \left| \int_{-\infty}^{+\infty} q'E_y dt \right| = \frac{qq'}{4\pi\epsilon_0} \int_{-\infty}^{+\infty} \frac{\gamma b}{(b^2 + \gamma^2 u^2 t^2)^{3/2}} dt = \frac{qq'}{4\pi\epsilon_0} \frac{2}{bu}. \quad (10.93)$$

Hence, the kinetic energy acquired by the scattering center (equal to the loss of energy of the incident particle) is

$$-\Delta\mathcal{E} = \frac{(\Delta p')^2}{2m'} = \left(\frac{qq'}{4\pi\epsilon_0} \right)^2 \frac{2}{m'u^2 b^2}. \quad (10.94)$$

Such energy losses have to be summed up over all collisions, with random values of the impact parameter b . At the scattering center density n , the number of collisions per small path length dz per small range db is $dN = n2\pi b db dz$, so that

$$-\frac{d\mathcal{E}}{dx} = \int (-\Delta\mathcal{E}) dN = n \left(\frac{qq'}{4\pi\epsilon_0} \right)^2 \frac{2}{m'u^2} 2\pi \int_{b_{\min}}^{b_{\max}} \frac{db}{b} = 4\pi n \left(\frac{qq'}{4\pi\epsilon_0} \right)^2 \frac{\ln B}{m'u^2}, \quad \text{where } B \equiv \frac{b_{\max}}{b_{\min}}. \quad (10.95)$$

Here the logarithmic integral over b was treated similarly to that over \mathcal{P} in the bremsstrahlung theory. This approach is adequate, because the ratio b_{\max}/b_{\min} is much larger than 1. Indeed, b_{\min} may be estimated from $(\Delta p')_{\max} \sim p = \gamma mu$. For this value, Eq. (93) with $q' \sim q$ gives $b_{\min} \sim r_c$ (see Eq. (8.41) and its discussion), which is, for elementary particles, of the order of 10^{-15} m. On the other hand, for the most important case when charges q' belong to electrons (which, according to Eq. (94) are the most efficient

³⁶ See, e.g., QM Secs. 6.3, 9.3, 9.5, and 9.7.

³⁷ See, e.g., MA Eq. (6.14).

³⁸ According to Eq. (9.139), $E_z = 0$, and the net impulse of the longitudinal force $q'E_x$ is zero.

Coulomb energy absorbers, due to their extremely low mass m' , b_{\max} may be estimated from condition $\tau = b/\gamma u \sim 1/\omega_{\max}$, where $\omega_{\max} \sim 10^{16} \text{ s}^{-1}$ is the characteristic frequency of electron transitions in atoms. (Below this frequency, our classical analysis of scatterer's motion is invalid.) From here, we have the estimate $b_{\max} \sim \gamma u/\omega_{\max}$, so that

$$B \equiv \frac{b_{\max}}{b_{\min}} \sim \frac{\gamma u}{r_c \omega_0}, \quad (10.96)$$

for $\gamma \sim 1$ and $u \sim c \approx 3 \times 10^8 \text{ m/s}$ giving $b_{\max} \sim 3 \times 10^{-8} \text{ m}$, and $B \sim 10^9$ (give or take a couple orders of magnitude – this does not change the estimate $\ln B \sim 20$ too much).³⁹

Now we can compare the Coulomb losses (95) with those due to the bremsstrahlung, given by Eq. (92):

$$\frac{-d\mathcal{E}|_{\text{radiation}}}{-d\mathcal{E}|_{\text{Coulomb}}} \sim \alpha \frac{m'}{m} \beta^2 \frac{1}{\ln B}, \quad (10.97)$$

Since $\alpha \sim 10^{-2} \ll 1$, for nonrelativistic particles ($\beta \ll 1$) the Coulomb losses of energy are much higher, and only for ultrarelativistic particles, the relation may be opposite.

According to Eq. (95), for electron-electron scattering ($q = q' = -e$, $m' = m_e$),⁴⁰ at the value $n \sim 6 \times 10^{26} \text{ m}^{-3}$ typical for air at ambient conditions, the characteristic length of energy loss,

$$l_c \equiv \frac{\mathcal{E}}{(-d\mathcal{E}/dx)}, \quad (10.98)$$

for electrons with kinetic energy $\mathcal{E} = 6 \text{ keV}$ is close to $2 \times 10^{-4} \text{ m} = 0.2 \text{ mm}$. (This is why you need vacuum in CRT monitors and electron microscope columns!) Since $l_c \propto \mathcal{E}^2$, more energetic particles penetrate deeper, until the bremsstrahlung steps in at very high energies.

10.5. Density effects and the Cherenkov radiation

For condensed matter, the Coulomb loss estimate made in the last section is not quite suitable, because it is based on the upper cutoff $b_{\max} \sim \gamma u/\omega_{\max}$. For the example given above, incoming electron velocity u is close to $5 \times 10^7 \text{ m/s}$, and for the typical value $\omega_{\max} \sim 10^{16} \text{ s}^{-1}$ ($\hbar\omega_{\max} \sim 10 \text{ eV}$), this cutoff $b_{\max} \sim 5 \times 10^{-9} \text{ m} = 5 \text{ nm}$. Even for air at ambient conditions, this is larger than the average distance ($\sim 2 \text{ nm}$) between the molecules, so that at the high end of the impact parameter range, at $b \sim b_{\max}$, the Coulomb loss events in adjacent molecules are not quite independent, and the theory needs corrections. For condensed matter, with much higher particle density n , most collisions satisfy condition

³⁹ A quantum analysis (carried out by H. Bethe in 1940) replaces, in Eq. (95), $\ln B$ with $\ln(2\gamma^2 m u^2 / \hbar \langle \omega \rangle) - \beta^2$, where $\langle \omega \rangle$ is the average frequency of the atomic quantum transitions weight by their oscillator strength. This refinement does not change the estimate given below. Note that both the classical and quantum formulas describe, a fast increase (as $1/\beta$) of the energy loss rate ($-d\mathcal{E}/dx$) at $\gamma \rightarrow 1$ and its slow increase (as $\ln \gamma$) at $\gamma \rightarrow \infty$, so that the losses have a minimum at $(\gamma - 1) \sim 1$.

⁴⁰ Actually, the above analysis has neglected the change of momentum of the incident particle. This is legitimate at $m' \ll m$, but for $m = m'$ the change approximately doubles the energy losses. Still, this does not change the order of magnitude of the estimate.

$$nb^3 \gg 1, \quad (10.99)$$

and the treatment of Coulomb collisions as independent events is completely inadequate. However, condition (99) enables the opposite approach: treating the medium as a continuum. In the time domain formulation, used in the previous sections of this chapter, this would be a very complex problem, because it would require an explicit description of medium dynamics. Here the frequency-domain approach, based on the Fourier transform in both time and space, helps a lot, provided that functions $\varepsilon(\omega)$ and $\mu(\omega)$ are considered known - either calculated or taken from experiment. Let us have a good look at such approach, because it gives some interesting (and practically important) results.

In Chapter 6, we have used the macroscopic Maxwell equations to derive Eqs. (6.109), which describe the time evolution of potentials in a medium with frequency-independent ε and μ . Looking for all functions participating in Eqs. (6.109) in the form of plane-wave expansion⁴¹

$$f(\mathbf{r}, t) = \int d^3k \int d\omega f_{\mathbf{k},\omega} e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)}, \quad (10.100)$$

and requiring all coefficients at similar exponents to be balanced, we get their Fourier image:⁴²

$$[k^2 - \omega^2 \varepsilon \mu] \phi_{\mathbf{k},\omega} = \frac{\rho_{\mathbf{k},\omega}}{\varepsilon}, \quad [k^2 - \omega^2 \varepsilon \mu] \mathbf{A}_{\mathbf{k},\omega} = \mu \mathbf{j}_{\mathbf{k},\omega}. \quad (10.101)$$

As was discussed in Chapter 7, in such a Fourier form, the Maxwell theory remain valid even for the dispersive media, so that Eq. (101) is generalized as

$$[k^2 - \omega^2 \varepsilon(\omega) \mu(\omega)] \phi_{\mathbf{k},\omega} = \frac{\rho_{\mathbf{k},\omega}}{\varepsilon(\omega)}, \quad [k^2 - \omega^2 \varepsilon(\omega) \mu(\omega)] \mathbf{A}_{\mathbf{k},\omega} = \mu(\omega) \mathbf{j}_{\mathbf{k},\omega}, \quad (10.102)$$

The evident advantage of these equations is that their formal solution is trivial:

$$\phi_{\mathbf{k},\omega} = \frac{\rho_{\mathbf{k},\omega}}{\varepsilon(\omega)[k^2 - \omega^2 \varepsilon(\omega) \mu(\omega)]}, \quad \mathbf{A}_{\mathbf{k},\omega} = \frac{\mu(\omega) \mathbf{j}_{\mathbf{k},\omega}}{[k^2 - \omega^2 \varepsilon(\omega) \mu(\omega)]}, \quad (10.103)$$

Field potentials in a linear medium

so that the “only” remaining things to do is to calculate the Fourier transforms of functions $\rho(\mathbf{r}, t)$ and $\mathbf{j}(\mathbf{r}, t)$, describing stand-alone charges and currents, using the transform reciprocal to Eq. (100), with one factor $1/2\pi$ per each scalar dimension,

$$f_{\mathbf{k},\omega} = \frac{1}{(2\pi)^4} \int d^3r \int dt f(\mathbf{r}, t) e^{-i(\mathbf{k} \cdot \mathbf{r} - \omega t)}, \quad (10.104)$$

and then carry out the integration (100).

For our current problem of a single charge q , uniformly moving in the medium with velocity \mathbf{u} ,

$$\rho(\mathbf{r}, t) = q \delta(\mathbf{r} - \mathbf{u}t), \quad \mathbf{j}(\mathbf{r}, t) = q \mathbf{u} \delta(\mathbf{r} - \mathbf{u}t), \quad (10.105)$$

the first task is easy:

⁴¹ All integrals here and below are in infinite limits, unless specified otherwise.

⁴² As was discussed in Sec. 7.2, the Ohmic conductivity of the medium (generally, also a function of frequency) may be readily incorporated into the dielectric permittivity: $\varepsilon(\omega) \rightarrow \varepsilon_{\text{ef}}(\omega) + i\sigma(\omega)/\omega$. In this section, I will assume that such incorporation, which is especially natural for high frequencies, has been performed, so that the current density $\mathbf{j}(\mathbf{r}, t)$ describes only stand-alone currents – for example, the current (105) of the incident particle.

$$\rho_{\mathbf{k},\omega} = \frac{q}{(2\pi)^4} \int d^3r \int dt q \delta(\mathbf{r} - \mathbf{u}t) e^{-i(\mathbf{k} \cdot \mathbf{r} - \omega t)} = \frac{q}{(2\pi)^4} \int e^{i(\omega t - \mathbf{k} \cdot \mathbf{u}t)} dt = \frac{q}{(2\pi)^3} \delta(\omega - \mathbf{k} \cdot \mathbf{u}). \quad (10.106)$$

Since expressions (105) for $\rho(\mathbf{r}, t)$ and $\mathbf{j}(\mathbf{r}, t)$ differ only by a constant factor \mathbf{u} , it is clear that the absolutely similar calculation for current would give

$$\mathbf{j}_{\mathbf{k},\omega} = \frac{q\mathbf{u}}{(2\pi)^3} \delta(\omega - \mathbf{k} \cdot \mathbf{u}). \quad (10.107)$$

Let us summarize what we have got by now, plugging Eqs. (106) and (107) into Eqs. (103):

$$\phi_{\mathbf{k},\omega} = \frac{1}{(2\pi)^3} \frac{q \delta(\omega - \mathbf{k} \cdot \mathbf{u})}{\varepsilon(\omega) [k^2 - \omega^2 \varepsilon(\omega) \mu(\omega)]}, \quad \mathbf{A}_{\mathbf{k},\omega} = \frac{1}{(2\pi)^3} \frac{\mu(\omega) q \mathbf{u} \delta(\omega - \mathbf{k} \cdot \mathbf{u})}{[k^2 - \omega^2 \varepsilon(\omega) \mu(\omega)]} = \varepsilon(\omega) \mu(\omega) \mathbf{u} \phi_{\mathbf{k},\omega}. \quad (10.108)$$

Now, at the last step of calculations, namely integration (100), we are starting to pay a heavy price for the easiness of the first steps. This is why let us think well what exactly do we need from it. First of all, for the calculation of power losses, the electric field is more convenient to use than the potentials, so let us calculate the Fourier images of \mathbf{E} and \mathbf{B} . Plugging expansion (100) into the fundamental relations (6.106), and again requiring the balance of exponent's coefficients, we get

$$\mathbf{E}_{\mathbf{k},\omega} = -i\mathbf{k}\phi_{\mathbf{k},\omega} + i\omega\mathbf{A}_{\mathbf{k},\omega} = i[\omega\varepsilon(\omega)\mu(\omega)\mathbf{u} - \mathbf{k}]\phi_{\mathbf{k},\omega}, \quad \mathbf{B}_{\mathbf{k},\omega} = i\mathbf{k} \times \mathbf{A}_{\mathbf{k},\omega} = i\varepsilon(\omega)\mu(\omega)\mathbf{k} \times \mathbf{u}\phi_{\mathbf{k},\omega}, \quad (10.109)$$

so that Eqs. (100) and (108) yield

$$\mathbf{E}(\mathbf{r}, t) = \int d^3k \int d\omega \mathbf{E}_{\mathbf{k},\omega} e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} = \frac{iq}{(2\pi)^3} \int d^3k \int d\omega \frac{[\omega\varepsilon(\omega)\mu(\omega)\mathbf{u} - \mathbf{k}] \delta(\omega - \mathbf{k} \cdot \mathbf{u})}{\varepsilon(\omega) [k^2 - \omega^2 \varepsilon(\omega) \mu(\omega)]} e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)}. \quad (10.110)$$

With the notation used in Eq. (51), this integral may be partitioned as

$$\mathbf{E}(\mathbf{r}, t) = \int \mathbf{E}_\omega e^{-i\omega t} d\omega, \quad \mathbf{E}_\omega = \int \mathbf{E}_{\mathbf{k},\omega} e^{i\mathbf{k} \cdot \mathbf{r}} d^3k = \frac{iq}{(2\pi)^3} \int \frac{[\omega\varepsilon(\omega)\mu(\omega)\mathbf{u} - \mathbf{k}] \delta(\omega - \mathbf{k} \cdot \mathbf{u})}{\varepsilon(\omega) [k^2 - \omega^2 \varepsilon(\omega) \mu(\omega)]} e^{i\mathbf{k} \cdot \mathbf{r}} d^3k. \quad (10.111)$$

Let us calculate the Cartesian components of the partial Fourier image \mathbf{E}_ω , at a point separated by distance b from particle's trajectory. Selecting the coordinates and time origin as shown in Fig. 9.11a, we have $\mathbf{r} = \{0, b, 0\}$, so that only E_x and E_y are not vanishing. In particular, according to Eq. (111),

$$(E_x)_\omega = \frac{iq}{(2\pi)^3 \varepsilon(\omega)} \int dk_x \int dk_y \int dk_z \frac{\omega\varepsilon(\omega)\mu(\omega)u - k_x}{k^2 - \omega^2 \varepsilon(\omega) \mu(\omega)} \delta(\omega - k_x u) e^{ik_y b}. \quad (10.112)$$

The delta-function kills one integral (over k_x) of three, and we get:

$$(E_x)_\omega = \frac{iq}{(2\pi)^3 \varepsilon(\omega) u} \left[\omega\varepsilon(\omega)\mu(\omega)u - \frac{\omega}{u} \right] \int e^{ik_y b} dk_y \int \frac{dk_z}{\omega^2 / u^2 + k_y^2 + k_z^2 - \omega^2 \varepsilon(\omega) \mu(\omega)}. \quad (10.113)$$

The last integral (over k_y) may be readily reduced to the table integral $\int d\xi / (1 + \xi^2)$, in infinite limits, equal to π .⁴³ The result may be presented as

⁴³ See, e.g., MA Eq. (6.5a).

$$(E_x)_\omega = -\frac{i\pi q\kappa^2}{(2\pi)^3 \omega \varepsilon(\omega)} \int \frac{e^{ik_y b}}{(k_y^2 + \kappa^2)^{1/2}} dk_y, \quad (10.114)$$

where parameter κ (generally, a complex function of frequency) is defined as

$$\kappa^2 \equiv \omega^2 \left(\frac{1}{u^2} - \varepsilon(\omega) \mu(\omega) \right). \quad (10.115)$$

The last integral may be expressed via the modified Bessel function of the second kind:⁴⁴

$$(E_x)_\omega = -\frac{iqu\kappa^2}{(2\pi)^2 \omega \varepsilon(\omega)} K_0(\kappa b). \quad (10.116)$$

A similar calculation yields

$$(E_y)_\omega = \frac{q\kappa}{(2\pi)^2 \varepsilon(\omega)} K_1(\kappa b). \quad (10.117)$$

Now, instead of rushing to make the final integration (111) over frequency to calculate $\mathbf{E}(t)$, let us realize that what we need for power losses is only the total energy loss through the whole time of particle passage. Energy loss per unit volume is

$$-\frac{d\mathcal{E}}{dV} = \int \mathbf{j} \cdot \mathbf{E} dt, \quad (10.118)$$

where \mathbf{j} is the current of bound charges in the medium, and should not be confused with the free particle's current (105). This integral may be readily expressed via the partial Fourier image \mathbf{E}_ω and the similarly defined image \mathbf{j}_ω , just as it was done at the derivation of Eq. (54):

$$-\frac{d\mathcal{E}}{dV} = \int dt \int d\omega e^{-i\omega t} \int d\omega' e^{-i\omega' t} \mathbf{j}_\omega \cdot \mathbf{E}_{\omega'} = 2\pi \int d\omega \int d\omega' \mathbf{j}_\omega \cdot \mathbf{E}_{\omega'} \delta(\omega + \omega') = 2\pi \int \mathbf{j}_\omega \cdot \mathbf{E}_{-\omega} d\omega. \quad (10.119)$$

In our approach, the Ohmic conductance is incorporated into the complex permittivity $\varepsilon(\omega)$, so that, according to the discussion in the end of Sec. 7.2, current's Fourier image is

$$\mathbf{j}_\omega = \sigma_{\text{ef}}(\omega) \mathbf{E}_\omega = -i\omega \varepsilon(\omega) \mathbf{E}_\omega. \quad (10.120)$$

As a result, Eq. (119) yields

$$-\frac{d\mathcal{E}}{dV} = -2\pi i \int \varepsilon(\omega) \mathbf{E}_\omega \cdot \mathbf{E}_{-\omega} \omega d\omega = 4\pi \text{Im} \int_0^\infty \varepsilon(\omega) |\mathbf{E}_\omega|^2 \omega d\omega. \quad (10.121)$$

(The last transition is possible due to the property $\varepsilon(-\omega) = \varepsilon^*(\omega)$, which was discussed in Sec. 7.2.)

Finally, just as in the last section, we have to calculate the energy loss rate averaged over random values of the impact parameter b :

⁴⁴ As a reminder, the main properties of these functions are listed in Sec. 2.5 of these notes – see, in particular, Fig. 2.20b and Eqs. (2.157)-(2.158).

$$-\frac{d\mathcal{E}}{dx} = \int \left(-\frac{d\mathcal{E}}{dV} \right) d^2b = 2\pi \int_{b_{\min}}^{\infty} \left(-\frac{d\mathcal{E}}{dV} \right) b db = 8\pi^2 \int_{b_{\min}}^{\infty} b db \int_0^{\infty} \left(|E_x|_{\omega}^2 + |E_y|_{\omega}^2 \right) \text{Im} \varepsilon(\omega) \omega d\omega. \quad (10.122)$$

Note that we are cutting the resulting integral over b from below at some b_{\min} where our theory loses legitimacy. (On that limit, we are not doing much better than in the past section). Plugging in the calculated expressions (116) and (117) for field components, swapping the integrals, and using recurrence relations (2.142), which are valid for any Bessel functions, we finally get:

Radiation
intensity

$$-\frac{d\mathcal{E}}{dx} = \frac{2}{\pi} q^2 \text{Im} \int_0^{\infty} (\kappa^* b_{\min}) K_1(\kappa^* b_{\min}) K_0(\kappa^* b_{\min}) \frac{d\omega}{\omega \varepsilon(\omega)}. \quad (10.123)$$

This general result is valid for an arbitrary linear medium, with arbitrary dispersion relations $\varepsilon(\omega)$ and $\mu(\omega)$. (The last function participates in Eq. (123) only via Eq. (115) which defines parameter κ .) To get more concrete results, some particular model of the medium should be used. Let us explore the Lorentz oscillator model, which was discussed in Sec. 7.2, in its form (7.33) suitable for transition to quantum-mechanical description of atoms:

$$\varepsilon(\omega) = \varepsilon_0 + \frac{nq'^2}{m} \sum_j \frac{f_j}{(\omega_j^2 - \omega^2) - 2i\omega\delta_j}, \quad \sum_j f_j = 1, \quad \mu(\omega) = \mu_0. \quad (10.124)$$

If the damping of the effective atomic oscillators is low, $\delta_j \ll \omega_j$, and particle's speed u is much lower than the typical wave's phase velocity v (and hence c !), then for most frequencies Eq. (115) gives

$$\kappa^2 \equiv \omega^2 \left(\frac{1}{u^2} - \frac{1}{v^2(\omega)} \right) \approx \frac{\omega^2}{u^2}, \quad (10.125)$$

i.e. $\kappa = \kappa^* \approx \omega/u$ is real. In this case, Eq. (123) may be shown to give Eq. (95) with

$$b_{\max} = \frac{1.123u}{\langle \omega \rangle}. \quad (10.126)$$

Good news here is that both approaches (the microscopic analysis of Sec. 4 and the macroscopic analysis of this section) give essentially the same result. This fact may be also perceived as bad news: the treatment of the medium as a continuum does not give any new results here. The situation somewhat changes at relativistic velocities at which such treatment provides noticeable corrections (called *density effects*), in particular reducing the energy loss estimates.

Let me, however, skip these details and focus on a much more important effect described by our formulas. Consider the dependence of the electric field components on the impact parameter b , i.e. on the closest distance between particle's trajectory and the field observation point. If $\kappa^2 > 0$, then κ is real, and we can use, in Eqs. (116)-(117), the asymptotic formula (2.158),

$$K_n(\xi) \rightarrow \left(\frac{\pi}{2\xi} \right)^{1/2} e^{-\xi}, \quad \text{at } \xi \rightarrow \infty, \quad (10.127)$$

to conclude that the complex amplitudes E_{ω} of both components E_x and E_y of the electric field decrease exponentially, starting from $b \sim u/\langle \omega \rangle$. However, let us consider what happens at frequencies where $\kappa^2 < 0$, i.e.

$$\varepsilon(\omega)\mu(\omega) \equiv \frac{1}{v^2(\omega)} < \frac{1}{u^2} < \frac{1}{c^2} = \varepsilon_0\mu_0. \quad (10.128)$$

(This condition means that particle's velocity is larger than the phase velocity of waves, at this particular frequency.) In these intervals, κ is purely imaginary,⁴⁵ functions $\exp\{\kappa b\}$ become just phase factors, and

$$|E_x(\omega)| \propto |E_y(\omega)| \propto \frac{1}{b^{1/2}}. \quad (10.129)$$

This means that the Poynting vector drops as $1/b$, so that its flux through a surface of a round cylinder of radius b , with the axis on the particle trajectory (i.e. power flow), does not depend on b . Hence, this is wave emission – the famous *Cherenkov radiation*.⁴⁶

The direction of its propagation may be readily found taking into account that at large distances from particle's trajectory the emitted wave has to be locally planar, so that the *Cherenkov angle* θ may be found from the ratio of the field components (Fig. 13a):

$$\tan \theta = -\frac{E_x}{E_y}. \quad (10.130)$$

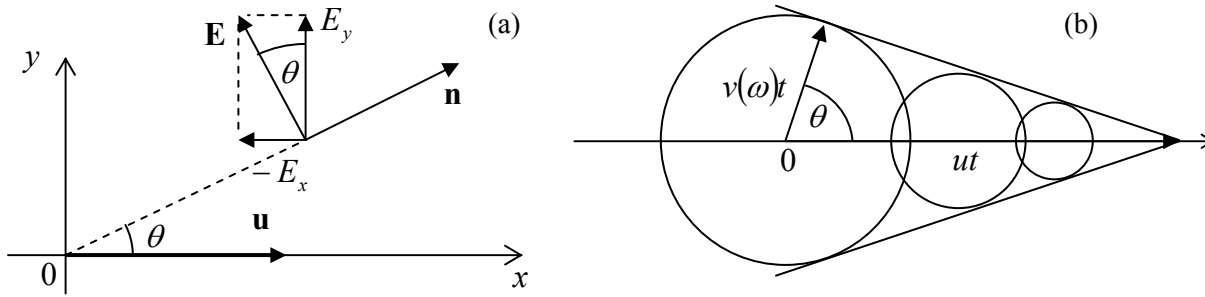


Fig. 10.13. (a) Cherenkov radiation's propagation angle θ , and (b) its interpretation.

This ratio may be calculated by plugging the asymptotic formula (127) into Eqs. (116) and (117) and calculating their ratio:

$$\tan \theta = -\frac{E_x}{E_y} = \frac{i\kappa u}{\omega} = [\varepsilon(\omega)\mu(\omega)u^2 - 1]^{1/2} = \left(\frac{u^2}{v^2(\omega)} - 1\right)^{1/2}, \quad (10.131a)$$

so that

⁴⁵ Strictly speaking, inequality $\kappa^2 < 0$ does not make sense for a medium with complex $\varepsilon(\omega)\mu(\omega)$ and hence complex $\kappa^2(\omega)$. However, in a typical medium where particles can propagate over substantial distances, the imaginary part of product $\varepsilon(\omega)\mu(\omega)$ does not vanish only in very limited frequency intervals, much more narrow than the intervals which we are now discussing - please have one more look at Fig. 7.5.

⁴⁶ This radiation was observed experimentally by P. Cherenkov (in older Western texts, “Čerenkov”) in 1934, with the observations explained by I. Frank and E. Tamm in 1937. Note, however, that the effect had been predicted theoretically as early as in 1889 by the same O. Heaviside whose name was mentioned so many times above - and whose genius I believe is still underappreciated.

Cherenkov
angle

$$\cos \theta = \frac{v(\omega)}{u} < 1. \quad (10.131b)$$

Remarkably, this direction does not depend on the emission time t' , so that radiation of frequency ω , at each instant, forms a hollow cone led by the particle. This simple result allows an evident interpretation (Fig. 13b): the cone is just the set of all observation points that may be reached by “signals” propagating with speed $v(\omega) < u$ from all previous points of particle’s trajectory.

This phenomenon is closely related to the so-called *Mach cone* in fluid dynamics,⁴⁷ besides that in the Cherenkov radiation there is a separate cone for each frequency (of the range in which $v(\omega) < u$): the smaller is the $\varepsilon(\omega)\mu(\omega)$ product, i.e. the larger is wave velocity $v(\omega) = 1/[\varepsilon(\omega)\mu(\omega)]^{1/2}$, and the broader is the cone, i.e. the earlier the corresponding “shock wave” arrives to an observer. Please note that the Cherenkov radiation is a unique radiative phenomenon: it takes place even if a particle moves without acceleration, and (in agreement with our analysis in Sec. 2), is impossible in free space where $v = c$ is always larger than u .

The intensity of the Cherenkov radiation intensity may be also readily found by plugging the asymptotic expression (127), with imaginary κ , into Eq. (123). The result is

Cherenkov
radiation
intensity

$$-\frac{d\mathcal{E}}{dx} \approx \left(\frac{ze}{4\pi}\right)^2 \int_{v(\omega) < u} \omega \left(1 - \frac{v^2(\omega)}{u^2}\right) d\omega. \quad (10.132)$$

For nonrelativistic particles ($u \ll c$), the Cherenkov radiation condition $u > v(\omega)$ may be fulfilled only in relatively narrow frequency intervals where the product $\varepsilon(\omega)\mu(\omega)$ is very large (usually, due to optical resonance peaks of the electric permittivity – see Fig. 7.5 and its discussion). In this case the emitted light consists of a few nearly monochromatic components. On the contrary, if the condition $u > v(\omega)$, i.e. $u^2/\varepsilon(\omega)\mu(\omega) > 1$ is fulfilled in a broad frequency range (as it is for ultrarelativistic particles in condensed media), the radiated power is clearly dominated by higher frequency of the range – hence the famous bluish color of the Cherenkov radiation glow in water nuclear reactors– see Fig. 14.

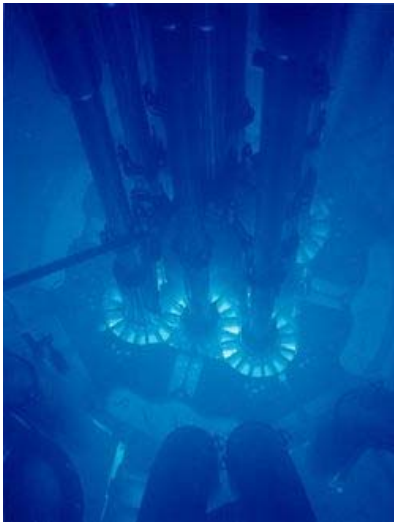


Fig. 10.14. Cherenkov radiation glow coming from the Advanced Test Reactor of the Idaho National Laboratory. Adapted from http://en.wikipedia.org/wiki/Cherenkov_radiation.

⁴⁷ See, e.g., a brief discussion in CM Sec. 8.6.

The Cherenkov radiation is broadly used for the detection of radiation in high energy experiments for particle identification and speed measurement (since it is easy to pass particles through media of various density and hence of the dielectric constant) – for example, in the so-called Ring Imaging Cherenkov (RICH) detectors that have been designed for the DELPHI experiment⁴⁸ at the Large Electron-Positron Collider (LEP) in CERN.

A little bit counter-intuitively, the formalism described in this section is also very useful for the description of an apparently rather different effect - the so-called *transition radiation* that takes place when a charged particle crosses a border between two media.⁴⁹ The effect may be understood as result of the time dependence of the electric dipole formed by the moving charge and its mirror image in the counterpart medium – see Fig. 15. In the nonrelativistic limit, the effect allows a straightforward description combining the electrostatics picture of Sec. 3.4 (see Fig. 3.9 and its discussion), and Eq. (8.27) - slightly corrected for polarization effects of the media. However, if particle's velocity u is comparable with the phase velocity of waves in either medium, the adequate theory of the transition radiation becomes very close to that of the Cherenkov radiation.

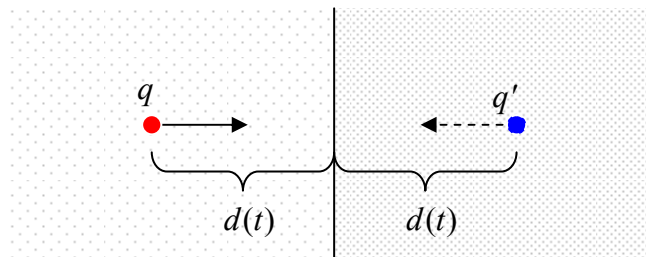


Fig. 10.15. Physics of the transition radiation.

In comparison with the Cherenkov radiation, the transition radiation is rather weak, and its practical use (mostly for the measurement of the relativistic factor γ , to which the radiation intensity is proportional) requires multi-layered stacks.⁵⁰ In these systems, the radiation emitted at sequential borders may be coherent, and the system's physics becomes close to that of the undulators discussed in Sec. 4.

10.6. Radiation's back-action

An attentive reader could notice that so far our treatment of charged particle dynamics has never been fully self-consistent. Indeed, in Sec. 9.6 we have analyzed particle's motion in various external fields, ignoring the fields radiated by particle itself, while in Sec. 8.2 and earlier in this chapter these fields have been calculated (admittedly, just for a few simple cases), but, again, their back-action on the emitting particle have been ignored. Only in few cases we have taken the back effects of the radiation

⁴⁸ See, e.g., <http://delphiwww.cern.ch/offline/physics/delphi-detector.html>. For a broader view at radiation detectors (including Cherenkov ones), the reader may be referred to the classical text by G. F. Knoll, *Radiation Detection and Measurement*, 4th ed., Wiley, 2010, and a newer treatment by K. Kleinknecht, *Detectors for Particle Radiation*, Cambridge U. Press, 1999.

⁴⁹ The effect was predicted theoretically in 1946 by V. Ginzburg and I. Frank, and only later observed experimentally.

⁵⁰ See, e.g., Sec. 5.3 in K. Kleinknecht's monograph cited above.

implicitly, via the energy conservation. However, even in these cases, the near-field components of the fields (such as the first term in Eq. (20a), that affect the moving particle most, have been ignored.

At the same time, it is clear that generally the interaction of a point charge with its own field cannot be always ignored. As the simplest example, if an electron is made to fly through a resonant cavity, thus inducing oscillations in it, and then is forced to return to it before the oscillations have decayed, its motion will be certainly affected by the oscillating fields, just as if they had been induced by another source. There is no conceptual problem with applying the Maxwell theory to such “field-particle rendezvous” effects; moreover, it is the basis of the engineering design of such electron devices as klystrons, magnetrons, and undulators.

A problem arises only when no finite “rendezvous” point is enforced by boundary conditions, so that the most important self-field effects are at $R \equiv |\mathbf{r} - \mathbf{r}'| \rightarrow 0$, the most evident example being the radiation of particle in free space, described earlier in this chapter. We already know that radiation takes away a part of charge’s kinetic energy, i.e. has to cause its deceleration. One should wonder, however, whether such self-action effects might be described in a more direct, non-perturbative way.

As the first attempt, let us try a phenomenological approach based on the already derived formulas for radiation power \mathcal{P} . For the sake of simplicity, let us consider a nonrelativistic point charge q in free space, so that \mathcal{P} is described by Eq. (8.27), with electric dipole moment’s derivative over time equal to $q\mathbf{u}$:

$$\mathcal{P} = \frac{Z_0 q^2}{6\pi c^2} \dot{\mathbf{u}}^2 = \frac{2}{3c^3} \frac{q^2}{4\pi\epsilon_0} \dot{\mathbf{u}}^2. \quad (10.133)$$

The most naïve approach would be to write the equation of particle’s motion in the form

$$m\dot{\mathbf{u}} = \mathbf{F}_{\text{ext}} + \mathbf{F}_{\text{self}}, \quad (10.134)$$

and try to calculate the radiation back-action force by requiring its instant power, $-\mathbf{F}_{\text{self}}\mathbf{u}$, to be equal to \mathcal{P} . However, with Eq. (133), this approach (say, for 1D motion) would give a very unnatural result,

$$F_{\text{self}} \propto \frac{\dot{u}^2}{u}, \quad (10.135)$$

that might diverge at some points of particle’s trajectory. This failure is clearly due to the retardation effect: as the reader may recall, Eq. (133) results from the analysis of radiation fields at *large* distances from the particle, e.g., from the second term in Eq. (20a), i.e. when the non-radiative first term (which is much larger at *small* distances, $R \rightarrow 0$) is ignored.

Before exploring the effects of this term, let us, however, make one more try with Eq. (133), considering its *average* effect on some periodic motion of the particle. To calculate the average, let us write

$$\overline{\dot{u}^2} \equiv \frac{1}{T} \int_0^T \dot{\mathbf{u}} \cdot \dot{\mathbf{u}} dt, \quad (10.136)$$

and integrate this identity, over the motion period, by parts:

$$\overline{\mathcal{P}} = \frac{2}{3c^3} \frac{q^2}{4\pi\epsilon_0} \overline{(\dot{\mathbf{u}})^2} = \frac{2}{3c^3} \frac{q^2}{4\pi\epsilon_0} \frac{1}{T} \left(\dot{\mathbf{u}} \cdot \mathbf{u} \Big|_0^T - \int_0^T \ddot{\mathbf{u}} \cdot \mathbf{u} dt \right) = -\frac{1}{T} \int_0^T \frac{2}{3c^3} \frac{q^2}{4\pi\epsilon_0} \ddot{\mathbf{u}} \cdot \mathbf{u} dt. \quad (10.137)$$

On the other hand, the back-action force would give

$$\overline{\mathcal{P}} = -\frac{1}{T} \int_0^T \mathbf{F}_{self} \cdot \mathbf{u} dt. \quad (10.138)$$

These two averages coincide if⁵¹

$$\mathbf{F}_{self} = \frac{2}{3c^3} \frac{q^2}{4\pi\epsilon_0} \ddot{\mathbf{u}}.$$

(10.139)

Abraham-
Lorentz
force

This is the so-called *Abraham-Lorentz force* for self-action. Before going after a more serious derivation of this formula, let us estimate its scale, presenting Eq. (139) as

$$\mathbf{F}_{self} = m \tau \ddot{\mathbf{u}}, \quad \text{with } \tau \equiv \frac{2}{3mc^3} \frac{q^2}{4\pi\epsilon_0}, \quad (10.140)$$

where constant τ evidently has the dimension of time. Recalling definition (8.41) of the classical radius r_c of the particle, Eq. (140) for τ may be rewritten as

$$\tau = \frac{2}{3} \frac{r_c}{c}. \quad (10.141)$$

For the electron, τ is of the order of 10^{-23} s. This means that in most cases the Abrahams-Lorentz force is either negligible or leads to the same results as the perturbative treatments of energy loss we have used earlier in this chapter.

However, Eq. (140) brings some unpleasant surprises. For example, let us consider a 1D oscillator of eigenfrequency ω_0 . For it, Eq. (134), with the back-action force given by Eq. (140), is

$$m\ddot{x} + m\omega_0^2 x = m\tau \ddot{x}. \quad (10.142)$$

Looking for the solution to this linear differential equation in the usual exponential form, $x(t) \propto \exp\{\lambda t\}$, we get the following characteristic equation,

$$\lambda^2 + \omega_0^2 = \tau \lambda^3. \quad (10.143)$$

It may look like that for any “reasonable” value of $\omega_0 \ll 1/\tau \sim 10^{23} \text{ s}^{-1}$, the right-hand side of this nonlinear algebraic equation may be treated as a perturbation. Indeed, looking for its solutions in the natural form $\lambda_{\pm} = \pm i\omega_0 + \lambda'$, with $|\lambda'| \ll \omega_0$, expanding both parts of Eq. (143) in the Taylor series in small parameter λ' , and keeping only linear terms, we get

⁵¹ This formula may be readily generalized to the relativistic case:

$$F_{self}^{\alpha} = \frac{2}{3mc^3} \frac{q^2}{4\pi\epsilon_0} \left[\frac{d^2 p^{\alpha}}{d\tau^2} + \frac{p^{\alpha}}{(mc)^2} \left(\frac{dp_{\beta}}{d\tau} \frac{dp^{\beta}}{d\tau} \right) \right],$$

- the so-called *Abraham-Lorentz-Dirac force*.

$$\lambda' \approx -\frac{\omega_0^2 \tau}{2}. \quad (10.144)$$

This means that the energy of free oscillations decreases in time as $\exp\{2\lambda't\} = \exp\{-\omega_0^2 \tau t\}$; this is exactly the radiative damping analyzed earlier. However, Eq. (143) is deceiving; it has the third root corresponding to unphysical, exponentially growing (so-called *run-away*) solutions. It is easiest to see for a free particle, with $\omega_0 = 0$. Then Eq. (143) becomes very simple,

$$\lambda^2 = \tau \lambda^3, \quad (10.145)$$

and it is easy to find all its 3 roots explicitly: $\lambda_1 = \lambda_2 = 0$ and $\lambda_3 = 1/\tau$. While the first 2 roots correspond to the values λ_{\pm} found earlier, the last one describes exponential (and extremely fast!) acceleration..

In order to remove this artifact, let us try to develop a self-consistent approach to back action, taking into account the near-field terms of particle fields. For that, we need somehow overcome the divergence of Eqs. (10) and (20) at $R \rightarrow 0$. The most reasonable way to do this is to spread particle charge over a ball of radius a , with a spherically-symmetric (but not necessarily constant) density $\rho(r)$, and in the end of calculations trace the limit $a \rightarrow 0$.⁵² Again sticking to the non-relativistic case (so that the magnetic component of the Lorentz force is not important), we should calculate

$$\mathbf{F}_{rad} = \int_V \rho(\mathbf{r}) \mathbf{E}(\mathbf{r}, t) d^3 r, \quad (10.146)$$

where the electric field is that of the charge itself, with field of any elementary charge $dq = \rho(r)d^3 r$, described by Eqs. (20a).

In order to make analytical calculations doable, we need to make assumption $a \ll r_c$, treat ratio $R/r_c \sim a/r_c$ as a small parameter, and expand the result in the Taylor series in small R . This procedure yields

$$\mathbf{F}_{self} = -\frac{2}{3} \frac{1}{4\pi\epsilon_0} \sum_{n=0}^{\infty} \frac{(-1)^n}{c^{n+2} n!} \frac{d^{n+1} \mathbf{u}}{dt^{n+1}} \int_V d^3 r \int_V d^3 r' \rho(r) R^{n-1} \rho(r'). \quad (10.147)$$

Distance R cancels only in the term with $n = 1$,

$$\mathbf{F}_1 = \frac{2}{3c^3} \frac{\ddot{\mathbf{u}}}{4\pi\epsilon_0} \int_V d^3 r \int_V d^3 r' \rho(r) \rho(r') = \frac{q^2}{6\pi\epsilon_0 c^3} \ddot{\mathbf{u}}, \quad (10.148)$$

showing that we have recovered (now in an *apparently* legitimate fashion) Eq. (139) for the Abrahams-Lorentz force. One could argue that in the limit $a \rightarrow 0$ the terms higher in $R \sim a$ (with $n > 1$) could be ignored. However, we have to notice that the main contribution to into series (147) is *not* described by Eq. (148) for $n = 1$, but is given by the larger term with $n = 0$:

$$\mathbf{F}_0 = -\frac{2}{3} \frac{1}{4\pi\epsilon_0} \frac{\dot{\mathbf{u}}}{c^2} \int_V d^3 r \int_V d^3 r' \frac{\rho(r) \rho(r')}{R} = -\frac{4}{3} \frac{\dot{\mathbf{u}}}{c^2} \frac{1}{8\pi\epsilon_0} \int_V d^3 r \int_V d^3 r' \frac{\rho(r) \rho(r')}{R} = -\frac{4}{3c^2} \dot{\mathbf{u}} U, \quad (10.149)$$

⁵² Note: this operation cannot be interpreted as describing a quantum spread due to the finite extent of point particle's wavefunction. In quantum mechanics, parts of wavefunction of the same charged particle do *not* interact with each other!

This term may be interpreted as the inertial “force” $-m_{\text{ef}}\mathbf{a}$ ⁵³ with the effective *electromagnetic mass*

$$m_{\text{ef}} = \frac{4}{3} \frac{U}{c^2}. \quad (10.150)$$

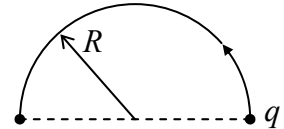
Effective
electro-
magnetic
mass

This is the famous (or rather infamous :-) *4/3 problem* that does not allow to interpret the electron’s mass as that of its electric field. The (admittedly, rather formal) resolution of this paradox is possible only in quantum electrodynamics with its renormalization techniques - beyond the framework of this course. Note that these issues are only important for motions with frequencies of the order of $1/\tau \sim 10^{23} \text{ s}^{-1}$, i.e. at energies $\mathcal{E} \sim \hbar/\tau \sim 10^{-11} \text{ J} \sim 10^8 \text{ eV}$, while other quantum electrodynamics effects may be observed at much lower frequencies, starting from $\sim 10^{10} \text{ s}^{-1}$. Hence the 4/3 problem is by no means the only motivation for the transfer from classical to quantum electrodynamics.

However, the reader should not think that his or her time spent on this course has been lost: quantum electrodynamics incorporates virtually all classical electrodynamics results, and transition between them is surprisingly straightforward.⁵⁴

10.6. Exercise problems

10.1. A point charge q that had been in a stationary position on a circle of radius R , is carried over, along the circle, to the opposite position on the same diameter (see Fig. on the right) as fast as only physically possible, and then is kept steady at this new position. Calculate and sketch the time dependence of the electric field \mathbf{E} at the center of the circle.



10.2. Express the total radiation power by a relativistic particle with the electric charge q and the rest mass m , moving with velocity \mathbf{u} , via the external Lorentz force \mathbf{F} exerted on the particle.

10.3. A relativistic particle with electric charge q , initially at rest, is accelerated by a constant force \mathbf{F} until it reaches certain velocity u , and then moves by inertia. Calculate the total energy radiated during the acceleration.

10.4.* Calculate the power spectrum of the radiation emitted by a relativistic particle with charge q , performing 1D harmonic oscillations with frequency ω and displacement amplitude a .

10.5. Analyze the polarization and the spectral contents of the synchrotron radiation in the direction propagating perpendicular to particle’s rotation plane. How do the results change if not one, but $N > 1$ similar particles move around the circle, at equal angular distances?

10.6. Calculate the time dependence of the kinetic energy of a charged relativistic particle performing synchrotron motion in a constant and uniform magnetic field \mathbf{B} , and hence emitting the synchrotron radiation. Sketch particle’s trajectory.

⁵³ See, e.g., CM Sec. 6.6.

⁵⁴ See, e.g., QM Chapter 9 and references therein.

Hint: You may assume that the energy loss is relatively slow ($-d\mathcal{E}/dt \ll \omega_c \mathcal{E}$), but should spell out the condition of validity of this assumption.

10.7. Find the polarization of the synchrotron radiation propagating within particle's rotation plane.

10.8. The basic quantum theory of radiation shows⁵⁵ that the electric dipole radiation by a particle is allowed only if its angular momentum change at the transition equals $\pm\hbar$.

(i) Estimate the change ΔL of the orbital momentum of an ultrarelativistic particle due to its emission of a single photon of the synchrotron radiation.

(ii) Does the quantum mechanics forbid such radiation? If not, why?

10.9. A relativistic particle moves along axis z , with velocity u_z , through an undulator - a system of permanent magnets providing (in the simplest model) a perpendicular magnetic field, whose distribution near axis z is sinusoidal:⁵⁶

$$\mathbf{B} = \mathbf{n}_y B_0 \cos k_0 z.$$

Assuming that the field is so weak that it causes only relatively small deviations of particle's trajectory from the straight line, calculate the angular distribution of the resulting radiation. What condition does this assumption impose on system's parameters?

10.10. Discuss possible effects of the interference of the undulator radiation from different periods of its static field distribution, in particular, calculate the angular positions of maxima of the radiation power density.

10.11. An electron, launched directly toward a plane surface of a perfect conductor, is instantly absorbed by it at the collision. Find the angular distribution and frequency spectrum of the electromagnetic waves radiated at this collision, if the initial kinetic energy T of the particle is much larger than conductor's workfunction ϕ . Give a semi-quantitative discussion of the limitations of your result.

10.12. A relativistic particle, with the rest mass m and electric charge q , flies with the velocity u by an immobile point charge q' , with the impact parameter b so large that the deviations of its trajectory from the straight line are negligible. Calculate the total energy loss due to the electromagnetic radiation during the passage. Formulate the conditions of validity of your result.

⁵⁵ See, e.g., EM Sec. 9.3, in particular Eq. (9.53) and its discussion.

⁵⁶ As the Maxwell equation for $\nabla \times \mathbf{H}$ shows, such field distribution cannot be created in any nonvanishing volume of free space. However, it may be created on a line – e.g., on particle's straight trajectory.

This page is
intentionally left
blank